UNIVERSITÉ LIBRE DE BRUXELLES

FUNCTIONAL GENOMICS

Pierard Florian

Professeur : V. Detours

Cours: BIOL-F-423

Master en Bioinformatique et Modélisation (MBM 1)

Année académique : 2015-2016

Table des matières

Analyse et tri des données

Les données du TCGA (The Cancer Genome Atlas) sont extraites avec le Firehose du Broad Institute. Nous avons analysé le cancer de la thyroïde (THCA) ainsi que l'âge de méthylation des échantillons grâce au programme R avec l'interface RStudio.

L'analyse d'expression des gènes à été réalisée grâce au programme GSEA (Gene Set Enrichment Analysis) ainsi que le package « samr » sur R.

Données du patient

Les données des patients contiennent l'ensemble des valeurs de chaque paramètre clinique pour chacun des patients et sont stockées dans le fichier « THCA.clin.merged.picked.txt » qui sera extrait dans la dataframe « patientsInfos » de notre script personnel.

Données de méthylation

L'ensemble des données de méthylation des patients à analyser est contenu dans le fichier « THCA-7.Rda » et ce dernier comprend à la fois les échantillons tumoraux et les échantillons normaux. Ces derniers doivent être séparés grâce au code-barres de l'échantillon. Ce code barre est formaté de manière à ce que le 4ème élément (séparé par des « - ») soit compris entre 01 et 09 en cas d'échantillon tumoral, entre 10 et 19 en cas d'échantillon normal et finalement entre 20 et 29 en cas d'échantillon contrôle, par exemple : TCGA-J8-A3O2-01A-11D-A23O-05 est un échantillon tumoral.

Les données tumorales de méthylation sont stockées dans la dataframe « tumorDNAm » tandis que celle des échantillons normaux sont stockées dans la dataframe « normalDNAm ». Nous possédons environ 350 échantillons tumoraux pour 50 échantillons normaux.

L'âge de méthylation est donné pour chaque échantillon, le code-barres est donc complet (exemple : TCGA-J8-A3O2-06A-11D-A23O-05) alors que les données du patient ne comptent que les 3 premiers éléments (exemple : TCGA-J8-A3O2). Dans certains cas, nous pouvons donc avoir un âge chronologique provenant d'un patient qui est associé à 2 âges de méthylation tumoraux (ce cas n'est pas arrivé dans les échantillons de patients normaux). Par exemple, pour le patient TCGA-J8-A3O2, nous avions l'âge de méthylation de TCGA-J8-A3O2-06A-11D-A23O-05 et de TCGA-J8-A3O2-01A-11D-A23O-05. Nous avons donc 2 échantillons de tumeur différents mais pour un même patient. Dans ce cas, nous avons décidé de traiter les données différemment en fonction de ce qui était demandé. Une explication du choix est fourni au début de chaque question.

Données de séquençage

Pour GSEA, les données de séquençage sont déjà normalisées et proviennent du fichier « THCA.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.data.txt ». Ces données ont donc été générées par RNAseq par la technique Illumina.

Pour SAMseq, les données de séquençage ne sont pas normalisées et proviennent des raw counts du fichier :

« THCA.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes__data.data.txt »

Tests statistiques

Pour toutes nos p-valeurs, nous avons utilisé ce code de significativité, marqué en jaune dans les tableaux des résultats :

0 =< * < 0.001 =< ** < 0.01 =< * < 0.05**

Question 1 : Corrélation entre l'âge chronologique et l'âge de méthylation

Pour cette question, le but étant de déterminer la véracité de l'estimation des âges de méthylation, j'ai décidé de garder les 2 données d'âge de méthylation tumoral pour un même patient car cela permet donc d'avoir une comparaison supplémentaire.

Au final, nous avons donc 350 échantillons tumoraux alors que nous ne possédons que 56 échantillons normaux.

L'âge chronologique des patients est retiré à partir des données des différents patients. Les données seront comparées grâce à un scatter plot dans lequel chaque point représente la valeur dans un groupe sur l'axe des X et la valeur dans l'autre groupe sur l'axe des Y. Si les 2 dataframes sont corrélés, les points devraient, dans notre cas, approximativement former une droite. Cette corrélation est analysée par la corrélation de Spearman.

Ce test non paramétrique mesure donc la dépendance entre 2 variables même si elles ne respectent pas une distribution normale. Cette dépendance peut être linéaire ou non car elle se base sur les rangs des différentes mesures au sein d'un même groupe. Cependant, elle doit respecter une fonction monotone, c'est-à-dire que lorsqu'une valeur varie dans un groupe, elle doit varier dans le même sens dans l'autre groupe. Si la corrélation est parfaite entre les 2 groupes, la corrélation de Spearman vaut +1 ou -1. Plus on s'approche de 0, moins les groupes sont corrélés. L'analyse de la p-valeur de cette corrélation nous permet de savoir si la corrélation est significative. Cependant, la p valeur fournie n'est pas exacte car nous possédons des exæquo dans nos données.

Age chronologique vs Age de méthylation tumoral

Dans ce cas, le but est de savoir si l'âge de méthylation d'une tumeur est une bonne estimation de l'âge chronologique du patient. Pour cela, nous comparons l'âge de méthylation d'une tumeur à l'âge chronologique du patient sur laquelle la tumeur a été prélevée.

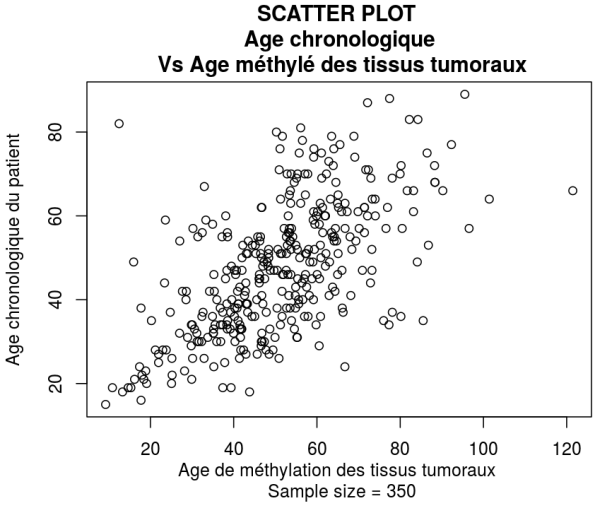


Illustration 1: Scatter plot de l'âge chronologique par rapport à l'âge de méthylation des tumeurs

La Corrélation de Spearman nous donne une valeur de 0,62 et la p-valeur calculée est de 3,112 e-38. Nous voyons donc que l'âge de méthylation des échantillons tumoraux est corrélé avec l'âge chronologique du patient.

Age chronologique vs Age de méthylation normal

Dans ce cas, le but est de savoir si l'âge de méthylation d'un tissu normal présente une meilleure estimation de l'âge chronologique du patient que l'âge de méthylation d'un tissu tumoral. Pour cela, nous comparons l'âge de méthylation d'un échantillon normal à l'âge chronologique du patient sur lequel l'échantillon a été prélevé.

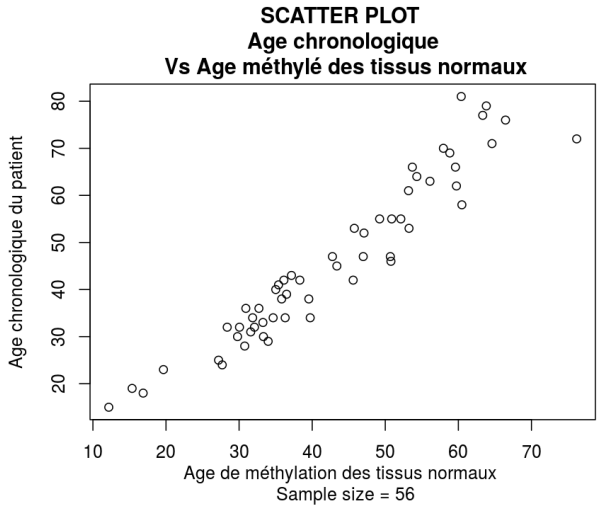


Illustration 2: Scatter plot de l'âge chronologique par rapport à l'âge de méthylation des tissus normaux

Dans ce cas, la corrélation de Spearman vaut 0,96. On observe donc une très bonne corrélation entre les 2 groupes car la valeur est proche de 1. La p-valeur calculée est de 5,460 e-33. L'âge de méthylation d'un tissu normal permet donc une estimation correcte de l'âge chronologique du patient.

Age de méthylation normal vs Age de méthylation tumoral

Cette comparaison permet de savoir si, pour un même patient, l'âge de méthylation d'une tumeur est corrélé à celui d'un tissu sain.

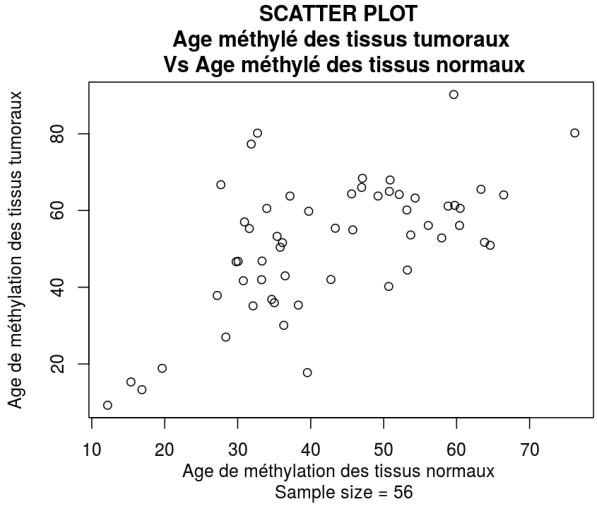


Illustration 3: Scatter plot de l'âge de méthylation des tissus normaux par rapport à l'âge de méthylation des tumeurs

Le scatter plot suffit à montrer la plus faible corrélation des données. La corrélation de Spearman vaut d'ailleurs 0,52 et la p-valeur calculée est de 5,786 e-05. Les données semblent sont donc corrélées mais moins que précédemment.

Accélération

Finalement, nous voulons déterminer si les 2 moyens de calcul de l'accélération de l'âge tumoral sont identiques. Tous 2 sont calculés à partir de l'âge de méthylation tumoral mais le premier est divisé par l'âge de méthylation du tissu normal alors que l'autre est divisé par l'âge réel du patient. Si ces 2 données sont corrélées, il devient inutile de mesurer l'âge de méthylation du tissu normal pour lequel nous avons moins de données.

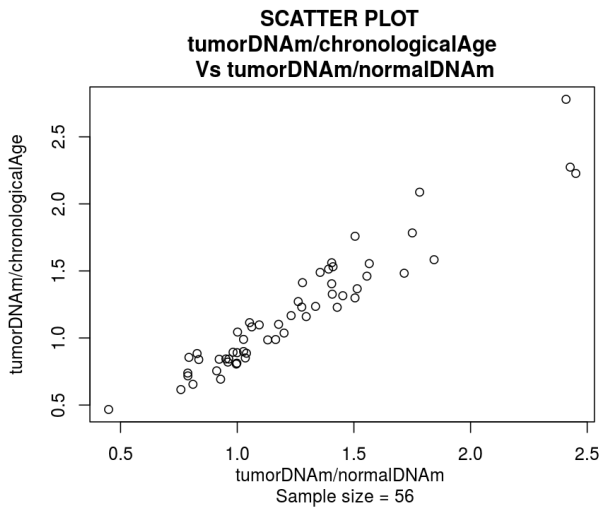


Illustration 4: Scatter plot de l'accélération en fonction de la façon de la calculer

La corrélation de Spearman étant de 0.95 avec une p-valeur de 0, nous pouvons conclure que l'accélération peut être calculée grâce au rapport entre l'âge de méthylation tumoral et l'âge chronologique du patient.

L'accélération représente donc le vieillissement plus rapide de l'estimation de l'âge de méthylation de la tumeur par rapport à l'age réel du patient. Par exemple, une accélération de 2 signifie que l'âge de méthylation tumoral est 2 fois plus élevé que l'âge réel du patient. La méthylation de l'ADN présentera donc des marqueurs d'un patient 2 fois plus vieux que son âge réel.

CONCLUSION QUESTION 1

Variable 1	Variable 2	Nombre de comparaisons	Corrélation de Spearman	**p-valeur**
TumorDNAm	Age chrono	350	0,62	**3,112 e-38**
NormalDNAm	Age chrono	56	0,96	**5,460 e-33**
NormalDNAm	TumorDNAm	56	0,52	**5,786 e-05**
TumorDNAm/ NormalDNAm	TumorDNAm/ Age chrono	56	0,95	**0**

Tableau 1: Conclusion de la question 1, comparaison des âges de méthylation, de l'âge réel ainsi que de l'accélération

Nous remarquons donc que l'âge de méthylation pour les tissus normaux semble parfaitement refléter l'âge chronologique du patient alors que l'âge de méthylation de la tumeur l'estime légèrement moins bien. Les marqueurs de méthylation de l'ADN permettant l'estimation

de l'âge réel du patient seraient donc modifiés en cas de tumeur. Il semble donc que les cellules cancéreuses vieillissent différemment des cellules normales au niveau de leur méthylation.

Question 2 : variables cliniques

Le but de cette partie est de déterminer si l'âge de méthylation dans les cancers est corrélé à différentes variables cliniques. Cette analyse est ensuite réalisée avec l'accélération à la place de l'âge de méthylation du cancer.

Pour cette question, lorsque nous avons 2 âges de méthylation pour un même patient, nous avons décidé de faire la moyenne de ces 2 valeurs. En effet, nous ne voulons pas augmenter le nombre de membres dans un groupe alors que c'est le même patient. Nous passons donc de 350 âges de méthylation tumoraux à 346.

Les différentes variables et les tests statistiques

Deux variables n'ont pas été analysées. Tout d'abord, l'âge du patient à déjà été analysé précédemment dans la question 1. Ensuite, le site de la tumeur ne sera également pas analysé car tous les échantillons proviennent de tumeurs de la thyroïde.

Afin d'analyser chaque variable, nous avons besoin de différents tests statistiques en fonction des variables. Tout d'abord, nous devons savoir quel type de données nous avons. Le test paramétrique est plus puissant que le test non paramétrique mais n'est applicable qu'en cas de distribution normale. Nous calculons également le nombre d'échantillons pour chaque groupe, car en cas de grand nombre, le test paramétrique peut être utilisé même si la normalité n'est pas totalement respectée.

Pour toutes ces analyses, les données NA n'ont logiquement pas été prise en compte.

- Les variables catégoriques où nous avons 2 groupes à différencier :
  - **Le sexe :** homme ou femme
  - **La radiothérapie :** oui ou non
  - **L'exposition à des radiations :** oui ou non
  - **La multifocalité :** unifocal ou multifocal. Décrit si la tumeur possède un ou plusieurs foyers.
  - **L'ethnicité :** « hispanique ou latino » ou « non hispanique ou latino »
  - **Le stade pathologique M:** décrit les métastases. m0 indique une absence de métastase alors que m1 en indique la présence. mx indique que cela n'a pas été évalué. Nous avons donc décidé de retirer mx des analyses.

Ensuite nous regardons la normalité de la distribution grâce à un histogramme :

- Distribution normale : nous utilisons un test paramétrique, le t-test avec la correction de Welch. Cette correction permet une meilleure analyse si les échantillons ont des variances ou des tailles d'échantillons différentes.
- Distribution non normale : nous utilisons un test non paramétrique, le test de Wilcoxon.

➢ <u>Les variables catégoriques où nous avons plus de 2 groupes :</u>

- **Le stade pathologique :** décrit la gravité de la tumeur ; stade 1, 2, 3, 4, 4a et 4c. Le stade le plus grave étant le stade 4c.
- **Le stade pathologique T :** décrit la taille de la tumeur et si elle est invasive dans les tissus adjacents ; t1, t1b, t1c, t2, t3 ou t4a. T0 est utilisé lorsque la tumeur primitive n'est pas localisée. Dans ce cas, elle se situe toujours au niveau de la thyroïde donc nous ne trouvons pas de T0. t4a représente les tumeurs les plus étendues.
- **Le stade pathologique N :** décrit les régions ganglionnaires proches touchées. N0 (absence), n1, n1a, n1b ou nx (non évalué).
- **Le type histologique :** classique/habituel, folliculaire, à grande cellule ou autre/non spécifié.
- **L'extension extrathyroïdienne :** non, minime ou avancée.
- **La tumeur résiduelle :** r0 (aucun), r1 (microscopique), r2 (macroscopique) et rx (non évalué)
- **La race :** blanc, noir/afro-américain, asiatique ou amérindien/natif d'Alaska

Ensuite, nous regardons la normalité de la distribution grâce à un boxplot :

- Distribution normale : nous utilisons un test paramétrique, la one-way ANOVA.
- Distribution non normale : nous utilisons un test non paramétrique, le test de Kruskal-Wallis.

- Les variables continues :

  - **Nombre de ganglions lymphatiques envahis**
  - **La taille de la tumeur**
  - **L'année de diagnostic**

  Nous réalisons ensuite un scatter plot dont on calcule la corrélation de Spearman ainsi que le test de corrélation de cette même méthode qui nous retourne une p-valeur afin de connaître la significativité.

- Les données de survie :

  - **Le statut vital :** vivant (0) ou mort (1)
  - **Le nombre de jour entre le diagnostic et la mort**
  - **Le nombre de jour entre le diagnostic et le dernier rendez-vous avec le médecin**

  Nous réalisons dans ce cas une étude de survie avec une courbe de Kaplan-Meier dont on calcule la régression de Cox.

Variables catégoriques en 2 groupes

Age de méthylation tumoral

Des histogrammes des 2 groupes pour chaque paramètre ont été générés afin de déterminer si les données respectaient une distribution normale ou non. Cela nous permet de choisir le test statistique adéquat.

Au vu des histogrammes (annexe 1), nous pouvons estimer que le genre, la radiothérapie et la multifocalité représentent une distribution normale en plus de présenter un grand nombre d'échantillons dans chaque groupe. Ces derniers seront donc analysés avec le t-test avec la correction de Welch (en noir dans le tableau 2).

Par contre les histogrammes de l'ethnicité, des expositions aux radiations et le stade pathologique M ne semblent pas présenter une distribution normale. De plus, un des 2 groupes présente chaque fois peu d'échantillons, et ce pour chacun des paramètres cliniques testés. Nous allons donc utiliser le test de Wilcoxon dans ces cas-ci (en rouge dans le tableau 2).

Paramètre clinique	Nombre d'échantillons groupe 1	Moyenne groupe 1	Nombre d'échantillons groupe 2	Moyenne groupe 2	**p-valeur**
Genre	87	50,8	259	51,5	**0,766**
Ethnicité	29	51,5	254	51,1	**0,753**
Radiothérapie	211	50,9	125	52,0	**0,586**
Expo. radiation	13	53,0	287	51,4	**0,723**
Multifocalité	157	51,9	186	50,9	**0,594**
Stade patho. M	7	58,9	199	51,6	**0,126**

Tableau 2: Analyse des variables à 2 groupes pour l'âge de méthylation tumoral

Aucun âge de méthylation des tumeurs n'est significativement différent entre les 2 groupes au sein d'un même paramètre clinique.

Accélération

Les même histogrammes sont générés afin de déterminer la normalité de la distribution (annexe 2). En ce qui concerne l'accélération, aucune des variables ne semble présenter une distribution normale évidente. Nous analysons donc nos données avec un test non paramétrique de Wilcoxon.

Paramètre clinique	Nombre d'échantillons groupe 1	Moyenne groupe 1	Nombre d'échantillons groupe 2	Moyenne groupe 2	**p-valeur**
Genre	87	1,06	259	1,14	**0,013 ***
Ethnicité	29	1,18	254	1,11	**0,270**
Radiothérapie	211	1,12	124	1,12	**0,602**
Expo. radiation	13	1,10	287	1,13	**0,645**
Multifocalité	157	1,12	186	1,12	**0,989**
Stade patho. M	7	1,13	199	1,11	**0,652**

Tableau 3: Analyse des variables à 2 groupes pour l'accélération

En comparaison avec les hommes, les femmes sembleraient donc avoir une accélération légèrement accentuée de l'âge de méthylation de leur tumeur par rapport à leur âge chronologique.

Variables catégoriques à plus de 2 groupes

Age de méthylation tumoral

Des boxplots ont été réalisés afin de comparer les différentes groupes au sein d'un même paramètre clinique (annexe 3).

Le stade pathologique N est le seul paramètre clinique pour lequel les différents groupes semblent suivre une distribution normale. Nous ajouterons donc une ANOVA one way pour ce dernier en plus du test de Kruskal-Wallis que nous avons réalisé pour tous les paramètres cliniques. L'ANOVA one way pour le stade pathologique N donne un résultat similaire avec une p-valeur de 0.114, n'étant donc pas significatif.

Paramètre clinique	Chi^2	Nombre d'échantillons total	Nombre de groupes	**p-valeur**
Ext. Extrahyroïd.	12,807	331	3	**0,002 ****
Stade patho.	55,251	344	6	**1,160 e-10 *****
Stade patho. T	26,381	346	8	**0,0004 *****
Stade patho. N	8,742	346	5	**0,069**
Race	1,981	289	4	**0,577**
Tumeur résiduelle	1,985	319	4	**0,576**
Type histo.	3,852	346	4	**0,278**

Tableau 4: Analyse des variables à plus de 2 groupes pour l'âge de méthylation tumoral

Les paramètres cliniques présentant une p-valeur significative (inférieur à 0,05) sont ensuite analysés grâce à un test post-hoc via la package « PMCMR » pour Pairwise Multiple Comparison of Mean Ranks.

Pour cela, le « posthoc.kruskal.nemenyi.test » est utilisé. Il permet de comparer les groupes les uns aux autres afin de savoir lesquels sont significativement différents.

Ext. Extrahyroïd.	Minimal	Moderate/advanced
None	0,036 *	0,011 *

Tableau 5: Test post hoc de l'extension extrathyroïdienne pour l'âge de méthylation tumoral

Nous observons donc que lorsqu'il n'y a pas d'extension extrathyroïdienne, l'âge de méthylation tumoral est significativement plus faible que lorsqu'il y a une extension, qu'elle soit

minime ou modérée. Cela peut s'expliquer par le fait que l'extension tumorale est une évolution du cancer de la thyroïde. Il paraît donc logique que l'âge de méthylation lorsqu'il y a une extension soit supérieur car le patient est probablement plus âgé.

Stade pathologique	Stade 2	Stade 3	Stade 4	Stade 4a	Stade 4c
Stade 1	0,0014 **	7,2 e-07 ***	0,97	2,2 e-05 ***	0,0361 *

Tableau 6: Test post hoc du stade pathologique pour l'âge de méthylation tumoral

Le stade pathologique 1 possède un âge de méthylation tumoral significativement inférieur aux autres stades (excepté le groupe 4 qui ne représente qu'une seule personne et n'est donc pas analysable). Cette observation peut-être expliquée de la même manière que pour l'extension extra thyroïdienne étant donné que les patients ayant un stade plus avancé de la maladie ont plus de chance d'être plus âgés.

Stade pathologique T	Stade t1	Stade t1a	Stade t1b	Stade t3	Stade t4	Stade t4a	Stade tx
Stade t2	0,0499 *	1	0,83	0,027 *	1	0,003 **	1

Tableau 7: Test post hoc du stade pathologique T pour l'âge de méthylation tumoral

Le stade pathologique t2 possède un âge de méthylation tumoral légèrement significativement plus bas que le stade t1 et t3. Par contre, le stade t2 possède une différence hautement significative avec le groupe t4a. Cette différence peut également être expliquée par l'âge plus important des patients ayant atteint le stade t4a.

Accélération

Les mêmes boxplots que ceux réalisés pour l'âge de méthylation ont été réalisés et sont présentés en annexe (annexe 4). Aucun paramètre clinique ne semble présenter tous ses groupes avec une distribution normale. Un test non paramétrique de Kruskall-Wallis est donc réalisé sur ces données afin de déterminer la présence d'un ou plusieurs groupes significativement différents des autres.

Paramètre clinique	Chi^2	Nombre d'échantillons total	Nombre de groupes	**p-valeur**
Ext. Extrahyroïd.	1,145	331	3	**0,564**
Stade patho.	28,108	344	6	**3,468 e-05 *****
Stade patho. T	13,044	346	8	**0,071**
Stade patho. N	11,259	346	5	**0,024 ***
Race	0,853	289	4	**0,837**
Tumeur résiduelle	3,186	319	4	**0,364**
Type histo.	6,204	346	4	**0,102**

Tableau 8: Analyse des variables à plus de 2 groupes pour l'accélération

Le « posthoc.kruskal.nemenyi.test » est ensuite réalisé afin de déterminer quel(s) groupe(s) se différencie(nt) des autres.

Stade pathologique	Stade 2	Stade 3	Stade 4	Stade 4a	Stade 4c
Stade 1	0,219	6,5 e-05 ***	0,557	0,055	0,975

Tableau 9: Test post hoc du stade pathologique pour l'accélération

Le stade pathologique 1 présente une accélération de l'âge de méthylation supérieure par rapport au stade 3 de manière très hautement significative. Au stade 1, la patient verrait son âge de méthylation tumoral augmenter plus rapidement que son âge réel par rapport au stade 3.

Stade pathologique N	Stade n0	Stade n1a	Stade n1b	Stade nx
Stade n1	0,099	0,159	0,084	0,012 *

Tableau 10: Test post hoc du stade pathologique N pour l'accélération

L'accélération est significativement différente entre le stade n1 et le stade nx (dont les données n'ont pas été évaluées). Cette information est donc inutile.

Variables continues

Age de méthylation tumoral

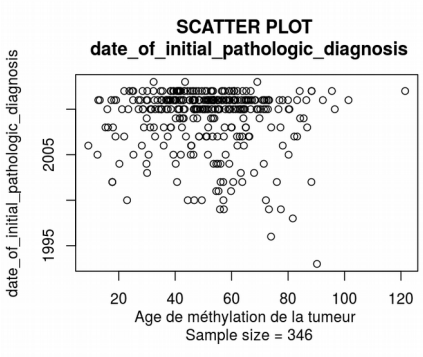


Illustration 5: Scatter plot de la date de diagnostic initial pour l'âge de méthylation tumoral

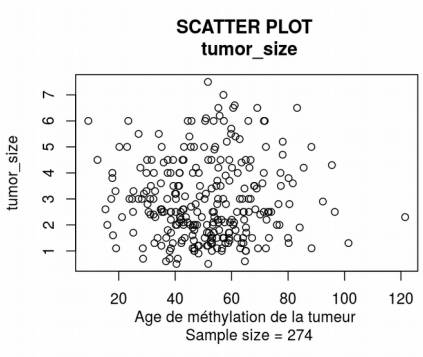


Illustration 6: Scatter plot de la taille de la tumeur pour l'âge de méthylation tumoral

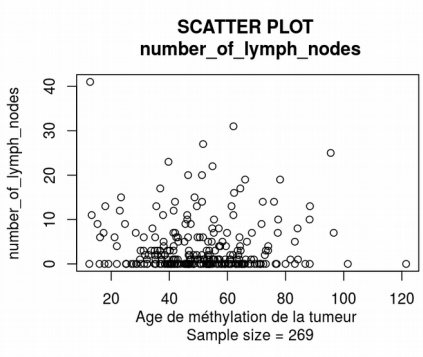


Illustration 7: Scatter plot du nombre de ganglions lymphatiques envahis pour l'âge de méthylation tumoral

Paramètres cliniques	Taille de l'échantillon	Corrélation de Spearman	**p-valeur**
Date de diagnostic	346	-0,06	**0,300**
Taille de la tumeur	274	0	**0,990**
Nombre de ggl. lymphatiques	269	-0,03	**0,584**

Tableau 11: Analyse des variables continues pour l'âge de méthylation tumoral

Accélération

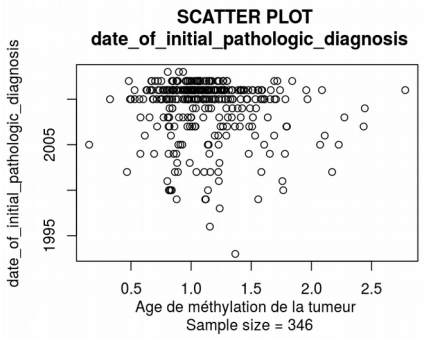


Illustration 8: Scatter plot de la date de diagnostic initial pour l'accélération

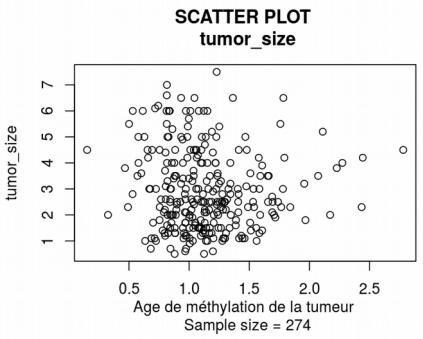


Illustration 9: Scatter plot de la taille de la tumeur pour l'accélération

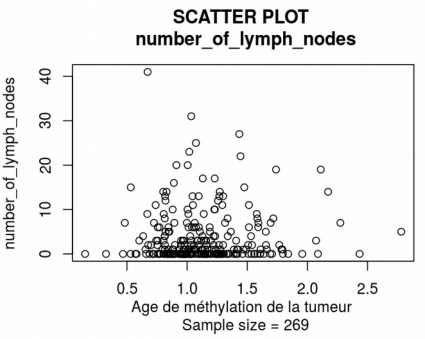


Illustration 10: Scatter plot du nombre de ganglions lymphatiques envahis pour l'accélération

Paramètres cliniques	Taille de l'échantillon	Corrélation de Spearman	**p-valeur**
Date de diagnostic	346	-0,09	**0,098**
Taille de la tumeur	274	-0,08	**0,166**
Nombre de gg. lymphatiques	269	0,04	**0,561**

Tableau 12: Analyse des variables continues pour l'accélération

Nous remarquons qu'aucune de ces variables cliniques ne présente une p-valeur inférieure à 0,05. De plus, le valeur absolue de la corrélation de Spearman est trop éloignée de 1. Les données ne semblent donc pas corrélées, que ce soit pour l'accélération ou pour l'âge de méthylation tumoral.

Données de survie

Le but ici est de déterminer si une variable a un impact sur le temps de survie. Une courbe de Kaplan-Meier permet donc la meilleure représentation de l'évolution de la survie au sein d'une étude. Pour cela, nous avons besoin de 3 informations de la dataframe patientsInfos :

Le « vital_status » qui représente l'occurrence de l'événement, ici le décès du patient. Si ce dernier vaut 1, le patient est décédé.

« days_to_death » représente donc le temps entre le diagnostic et le décès.

« days_to_last_followup » représente le temps entre le diagnostic et le dernier rendez vous. Après ce dernier rendez-vous, nous n'avons plus d'information sur le statut vital du patient. Nous ne pouvons donc pas être sûrs qu'il est en vie. Cet élément permet donc l'observation des événements de censure, représentés par une ligne verticale sur la courbe de Kaplan-Meier (Illustration 11).

Pour cela, un objet de survie doit donc être créé dans R grâce au package « survival ». Les « days_to_death » et « days_to_last_followup » sont donc rassemblés en 1 seule colonne car ils sont mutuellement exclusifs. Une 2ème colonne contenant l'événement de survie est ensuite créée. L'association de ces 2 colonnes permet la création de l'objet de survie, appelé « SurvObj » dans notre script. Ce dernier permet de générer une courbe de Kaplan-Meier avec les événements de censure.

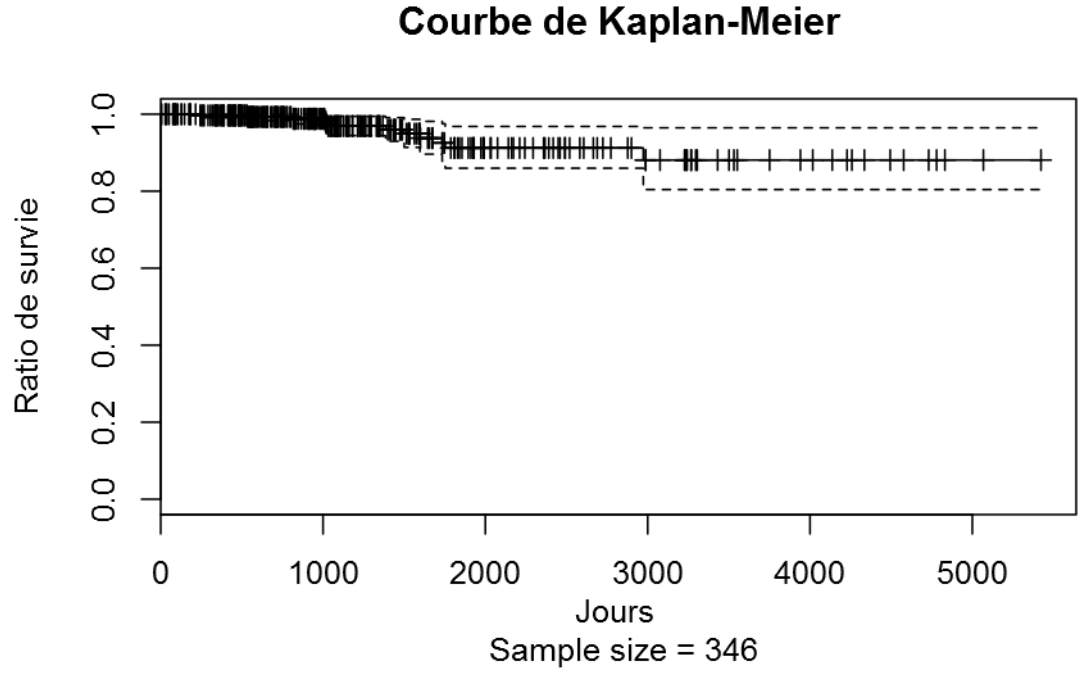


Illustration 11: Courbe de Kaplan-Meier pour l'analyse de la survie

Comme nous pouvons l'observer, très peu de patients décèdent durant l'étude et de nombreux patients ne sont plus suivis assez rapidement après le diagnostic (données censurées).

La comparaison de cet objet de survie avec le régression de Cox permet de déterminer si l'âge de méthylation de la tumeur est corrélé à la survie du patient au cours du temps.

Variables	Taille de l'échantillon	Nombre de décès	**Pr(>|z|)**
Age de méthylation tumoral	346	12	**0,004 ****
Accélération	346	12	**0,024 ***

Tableau 13: Analyse des données de survie

Il est logique que l'âge de méthylation tumoral soit significativement corrélé avec la survie étant donné qu'il est corrélé à l'âge chronologique. Dès lors, un patient plus âgé a normalement plus de risque de décéder.

La corrélation entre l'accélération et les données de survie est légèrement significative.

Très peu de patients sont décédés au cours de l'étude (12 sur 346) et de nombreux patients n'ont pas été suivi longtemps après leur diagnostic. Il convient dès lors de prendre ces résultats avec beaucoup de précautions.

CONCLUSION QUESTION 2

Durant ces nombreux tests, quelques corrélations sont ressorties comme significatives.

Étant donné la corrélation entre l'âge chronologique du patient et l'âge de méthylation tumoral (démontré dans la question 1), les variables pouvant être influencées par l'âge ont donc beaucoup de chances de ressortir positives lorsque nous comparons les variables à l'âge de méthylation tumoral. En effet, le cancer est une maladie évoluant au fil de temps, les patients âgés ayant donc plus de chances à la fois d'être à des stades pathologiques (T,N,M) plus avancés ou d'avoir une extension extrathyroïdienne mais également d'avoir une moins bonne probabilité de survie .

L'analyse de l'accélération représente le ratio entre l'âge de méthylation tumoral et l'âge chronologique du patient. Ici, les résultats significatifs ne peuvent donc pas s'expliquer par des stades plus avancés de la maladie chez les patients plus âgés.

Étonnement, les femmes présentent une accélération légèrement plus importante que les hommes. Les données de survie indiquent également une légère corrélation entre la probabilité de survie et l'accélération. Il est possible que les marqueurs de méthylation utilisés pour déterminer l'âge de méthylation soient également impliqués dans les probabilités de survie du patient. Pour s'en assurer, il faudrait regarder les régions utilisées pour déterminer l'âge de méthylation et voir si elles sont également impliquées dans la survie du patient. Le stade pathologique 1 présente une accélération de l'âge de méthylation significativement plus importante par rapport au stade 3. Il est donc possible que la tumeur voie son âge de méthylation augmenter plus rapidement que l'âge chronologique du patient principalement au début de l'évolution de la tumeur.

Ces analyses sont toujours à prendre avec précaution. Elles ne fournissent qu'une corrélation éventuelle et aucune causalité ne peut donc être déduite bien qu'elle puisse être supposée.

Question 3 : gene expression

Pour cette question, nous avons décidé de garder les données des 2 échantillons si nous avions 2 échantillons de tumeur pour un même patient. En effet, notre but ici est de comparer la variable à l'échantillon, il est donc préférable de garder un échantillon de plus à analyser.

GSEA

GSEA, pour Gene Set Enrichment Analysis, est un programme permettant de calculer si des sets de gènes sont exprimés significativement différemment par rapport à une variable. Ici la variable utilisée est continue à la fois pour l'âge de méthylation tumoral et l'accélération.

Nous avons décidé d'utiliser l'application java de GSEA qui fournit donc une interface graphique. Les données utilisées pour le séquençage ont déjà été normalisées et proviennent du fichier :
« THCA.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.data.txt ».

Prétraitement des données

GSEA demande un formatage des données très précis pour fonctionner, plusieurs fichiers ont donc du être créés, tous ayant leur propre formatage :

- SeqDatasClearedIDwith0.gct : Ce fichier contient les données de séquençage ayant subi un nettoyage. Pour cela, nous avons:
  - ✔ Supprimé les gènes dont le nom était inconnu, marqué par un « ? ». Nous avons ensuite gardé l'élément après le « | » représentant l'ID du gène.
  - ✔ Supprimé les colonnes dont l'échantillon ne correspondait à aucun échantillon dont la variable était connue.
  - ✔ Transformé les données en log2. Soit x l'expression d'un gène, nous avons appliqué la formule log2(x+1) afin de ne pas avoir des valeurs infinies lorsque l'expression était nulle.
  - ✔ Une 2ème colonne a été également ajoutée et complétée par des « NA » afin de correspondre au formatage attendu.
  - ✔ La première ligne doit valoir « #1.2 » et la seconde ligne doit contenir le nombre de probes (= lignes) suivi du nombre d'échantillons (= colonnes – 2).

#1.2						
20512	349					
NAME	DESCRIPTION	TCGA-4C-A93U-01A	TCGA-BJ-A0Z2-01A	TCGA-BJ-A0Z5-01A	TCGA-BJ-A0Z9-01A	TCGA-BJ-A0ZH-01A
57714	NA	9.24507038045918	8.06617047434371	7.4654431480868	9.19975183808729	7.77617506801209
645851	NA	3.27734506027868	4.35612267684235	2.27744921470538	1.99729240763651	2.77827166663844
652919	NA	5.10153662578659	5.91656501874501	5.3964438068201	0	4.55671755780317
653553	NA	7.82402888086361	8.18398565457983	8.11708981070874	6.8102104337992	8.87915405292075

Illustration 12: Exemple du fichier SeqDatasClearedIDwith0.gct, pour GSEA

- Phenotype.cls : Ce fichier contient les variables à analyser ainsi que le style des variables (numeric). Ici, nous utilisons des variables numériques et les variables sont l'âge de méthylation tumoral ou l'accélération. Toutes les données sont ensuite séparées par des tabs.

#numeric					
#TumorDNAm					
69.0975653690474	80.4735828488247	59.4925217053658	76.5887187878289	50.6204870424728	46.6089298007809
#Acceleration					
0.933750883365505	1.41181724296184	1.02573313285113	1.34366173311981	0.973470904662939	1.607204475889

Illustration 13: Exemple du fichier Phenotype.cls, pour GSEA

- OurDatas.chip : Ce fichier contient le gène associé à l'ID ainsi qu'une colonne titre contenant une description du gène, ici « NA ».

Probe Set ID	Gene Symbol	Gene Title
1	A1BG	NA
54715	A2BP1	NA
87769	A2LD1	NA
144568	A2ML1	NA

Illustration 14: Exemple du fichierOurDatas.chip, pour GSEA

- c2.cp.v5.1.entrez.gmt : Ce fichier est exporté de la MsigDB. Il contient les ID des gènes en fonction des pathways dans lesquels ils sont impliqués. Nous avons choisi c2.cp pour les curated genes des voies canoniques. Ce fichier n'est pas indispensable car nous choisissons « collpase dataset to gene symbols » = false.

KEGG_GLYCOLYSIS_GLUCONEOGENESIS	http://www.broadinstitute.	55902	2645	5232	5230	5162	5160
KEGG_CITRATE_CYCLE_TCA_CYCLE	http://www.broadinstitute.o	3420	1743	5106	1431	5162	5105
KEGG_PENTOSE_PHOSPHATE_PATHWAY	http://www.broadinstitute.	6120	22934	55276	25796	5634	8789

Illustration 15: Exemple du fichier c2.cp.v5.1.entrez.gmt, pour GSEA

GSEA est ensuite lancé avec comme option « collapse dataset to gene symbols » = *false* ainsi que le « metric for ranking genes » = *Pearson*. L'expression dataset est *SeqDatasClearedIDwith0.gct* , le gene sets database est *c2.cp.v5.1.entrez.gmt* , le phénotype labels est une fois *Phenotype.cls#TumorDNAm* et une fois *Phenotype.cls#Acceleration* et finalement le chip platform est *ourDatas.chip*. Les autres options sont laissées par défaut.

Age de méthylation tumoral

Corrélation positive

- 356 / 1071 gene sets are upregulated in phenotype **TumorDNAm_pos**
- 0 gene sets are significant at FDR < 25%
- 1 gene sets are significantly enriched at nominal pvalue < 1%
- 3 gene sets are significantly enriched at nominal pvalue < 5%

GENE SET (Corrélation positive)	Size	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
REACTOME_SULFUR_AMINO_ACID_METABOLISM	24	0.58	1.80	0.006	1.0	0.639	2783
REACTOME_ENERGY_DEPENDENT_REGULATION_OF_MTOR_BY_LKB1_AMPK	17	0.59	1.67	0.013	1.0	0.908	5136
KEGG_BIOSYNTHESIS_OF_UNSATURATED_FATTY_ACIDS	22	0.49	1.59	0.025	1.0	0.975	3672
REACTOME_PEROXISOMAL_LIPID_METABOLISM	20	0.56	1.54	0.077	1.0	0.988	2158
REACTOME_SPHINGOLIPID_DE_NOVO_BIOSYNTHESIS	30	0.49	1.51	0.065	1.0	0.994	4438

Tableau 14: Top 5 des pathways corrélés positivement avec l'âge de méthylation tumoral. Le classement est réalisé par ordre décroissant de Normalized Enrichment Score (NES)

Corrélation négative

- 715 / 1071 gene sets are upregulated in phenotype **TumorDNAm_neg**
- 0 gene sets are significantly enriched at FDR < 25%
- 2 gene sets are significantly enriched at nominal pvalue < 1%
- 12 gene sets are significantly enriched at nominal pvalue < 5%

GENE SET (Corrélation négative)	Size	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
BIOCARTA_MITOCHONDRIA_PATHWAY	21	-0.63	-2.00	0.000	0.415	0.144	5821
BIOCARTA_CASPASE_PATHWAY	23	-0.67	-1.75	0.008	1.000	0.733	4385
REACTOME_INTRINSIC_PATHWAY_FOR_APOPTOSIS	29	-0.52	-1.72	0.010	1.000	0.800	5664
REACTOME_DEPOSITION_OF_NEW_CENPA_CONTAINING_NUCLEOSOMES_AT_THE_CENTROMERE	60	-0.52	-1.70	0.034	1.000	0.838	5080
REACTOME_PHOSPHORYLATION_OF_THE_APC_C	17	-0.59	-1.68	0.027	1.000	0.872	3128

Tableau 15: Top 5 des pathways corrélés négativement avec l'âge de méthylation tumoral. Le classement est réalisé par ordre croissant de Normalized Enrichment Score (NES)

Accélération

Corrélation positive

- 662 / 1071 gene sets are upregulated in phenotype **Acceleration_pos**
- 0 gene sets are significant at FDR < 25%
- 1 gene sets are significantly enriched at nominal pvalue < 1%
- 11 gene sets are significantly enriched at nominal pvalue < 5%

GENE SET (Corrélation poitive)	Size	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
REACTOME_CELL_CELL_JUNCTION_ORGANIZATION	56	0.54	1.75	0.016	1.000	0.803	2453
REACTOME_TIGHT_JUNCTION_INTERACTIONS	29	0.60	1.73	0.019	1.000	0.848	3526
REACTOME_PRE_NOTCH_TRANSCRIPTION_AND_TRANSLATION	27	0.63	1.68	0.017	1.000	0.909	4631
PID_NECTIN_PATHWAY	30	0.60	1.63	0.035	1.000	0.948	5101
REACTOME_SIGNALING_BY_HIPPO	20	0.65	1.60	0.042	1.000	0.962	5393

Tableau 16: Top 5 des pathways corrélés positivement avec l'accélération. Le classement est réalisé par ordre décroissant de Normalized Enrichment Score (NES)

Corrélation négative

- 409 / 1071 gene sets are upregulated in phenotype **Acceleration_neg**
- 0 gene sets are significantly enriched at FDR < 25%
- 0 gene sets are significantly enriched at nominal pvalue < 1%
- 4 gene sets are significantly enriched at nominal pvalue < 5%

GENE SET (Corrélation négative)	Size	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX
REACTOME_MITOCHONDRIAL_TRNA_AMINOACYLATION	21	-0.64	-1.73	0.035	1.000	0.791	3350
KEGG_CARDIAC_MUSCLE_CONTRACTION	73	-0.49	-1.65	0.037	1.000	0.904	4158
REACTOME_STEROID_HORMONES	29	-0.48	-1.61	0.020	1.000	0.949	6854
KEGG_PARKINSONS_DISEASE	113	-0.62	-1.60	0.087	1.000	0.952	4704
KEGG_PORPHYRIN_AND_CHLOROPHYLL_METABOLISM	40	-0.49	-1.59	0.041	1.000	0.960	7581

Tableau 17: Top 5 des pathways corrélés négativement avec l'accélération. Le classement est réalisé par ordre croissant de Normalized Enrichment Score (NES)

Conclusion

GSEA a l’avantage de renvoyer des pathways ainsi que le nombre de gènes de ce pathway qui sont corrélés avec la variable.

A la fois l’âge de méthylation tumoral et l’accélération possèdent des p-valeurs qui sont inférieures à 0,05. Cependant, aucun de nos résultats ne possède un False Discovery Rate (FDR) inférieur à 25 %. Cela veut dire que dans plus de 25 % des permutations, ce gene set est resté positif uniquement par chance.

Aucun gene set n’a donc pu être retiré comme significativement différent tout en ayant un FDR inférieur à 25 %. Nous n’avons pas pu démontrer une quelconque corrélation entre l’expression d’un gene set et l’âge de méthylation tumoral ou l’accélération.

SAM

Le logiciel est fait à la base pour traiter les « raw counts ». Nous avons donc retirer les informations du fichier avec les données de séquençage non normalisées :

« THCA.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes__data.data.txt »

Toutes les analyses se font sur R via la package « samr ».

Prétraitement des données

SAMseq est un programme conçu pour l'analyse des RNAseq. Ce programme nécessite un formatage légèrement différent que pour GSEA. Nous avons besoin de 2 types de données.

- Matrice de séquençage : Les données étant extraites du fichier non normalisé, ces dernières nécessitent un nettoyage plus important :
  - ✔ Tout d'abord, nous supprimons les colonnes « scaled_estimate » et « transcript_id » afin de ne garder que les « raw_count » ainsi que la colonne « gene_id ».
  - ✔ Pour la colonne gene_id, nous gardons le GENE SYMBOL se situant avant le « | ». Attention, « SLC35E2 » se trouve en 2 exemplaires. Nous avons donc modifié le 2ème en mettant « SLC35E2_probe2 ». Nous avons ensuite attribué cette colonne à rownames.
  - ✔ Nous avons ensuite sélectionné les données des patients dont nous avions l'âge de méthylation et l'accélération et trié par ordre alphabétique les patients.
  - ✔ Aucune transformation logarithmique n'a été réalisée et les données ont été arrondies car le programme trouvait des valeurs qui n'étaient pas des integers.
  - ✔ Aucune normalisation n'est réalisée car le SAMseq semble la réaliser lui-même.
  - ✔ Finalement, la dataframe a été transformée en matrice.

	TCGA-4C-A93U-01A	TCGA-BJ-A0Z2-01A	TCGA-BJ-A0Z5-01A	TCGA-BJ-A0Z9-01A	TCGA-BJ-A0ZH-01A	TCGA-BJ-A18Y-01A	TCGA-BJ-A18Z-01A	TCGA-BJ-A190-01A
A1BG	225	714	451	205	578	319	340	121
A1CF	0	0	0	0	0	0	0	0
A2BP1	0	53	12	2	13	80	8	1
A2LD1	166	218	169	205	327	163	181	152

Illustration 16: Exemple de matrice de séquençage, pour SAM

- Vecteur de la variable : Il s'agit juste d'un vecteur numérique comprenant une fois les âges de méthylation tumoraux des différents patients et une fois l'accélération. Ce vecteur doit donc respecter l'ordre des échantillons. Pour cela, nous trions toujours nos données par ordre alphabétique, évitant ainsi tout problème.

SAMseq est ensuite lancé avec 100 permutions et en précisant que nos données sont quantitatives :

```
samfit = SAMseq(matrice, vecteur, resp.type = "Quantitative" , nperms = 100, genenames = rownames(matrice))
```

Les données des gènes surexprimés ou sous-exprimés sont fournies grâce à :

```
tableVariable = samfit$siggenes.table
```

Ensuite, les gènes surexprimés sont retirés grâce à la fonction :

```
tableVariable$genes.up
```

Finalement, les gènes sous exprimés sont retirés grâce à la fonction :

```
tableVariable$genes.lo
```

Nous avons présenté les 10 gènes sur ou sous exprimés, chaque fois ayant la plus faible q-value, représentant ainsi la plus faible probabilité d’avoir des faux positifs.

Age de méthylation tumoral

UP RÉGULÉ		
Gene Symbol	Score(d)	**q-value(%)**
MANEAL	0.32	**0**
ALOXE3	0.261	**0**
FAM150A	0.257	**0**
ETNK2	0.256	**0**
LHFPL4	0.251	**0**
C1orf115	0.25	**0**
TOMM20	0.24	**0**
TMEM25	0.24	**0**
RGMB	0.239	**0**
ANP32E	0.239	**0**

DOWN RÉGULÉ		
Gene Symbol	Score(d)	**q-value(%)**
RANBP17	-0.392	**0**
ZNF518B	-0.36	**0**
MSL3L2	-0.334	**0**
C1orf59	-0.305	**0**
POMC	-0.297	**0**
CABYR	-0.289	**0**
RFPL1S	-0.289	**0**
KCNS1	-0.288	**0**
C7orf13	-0.282	**0**
CCT5	-0.276	**0**

Tableau 18: Top 10 des gènes up-régulés et down-régulés pour l'âge de méthylation tumoral. Analyse réalisée avec SAMseq

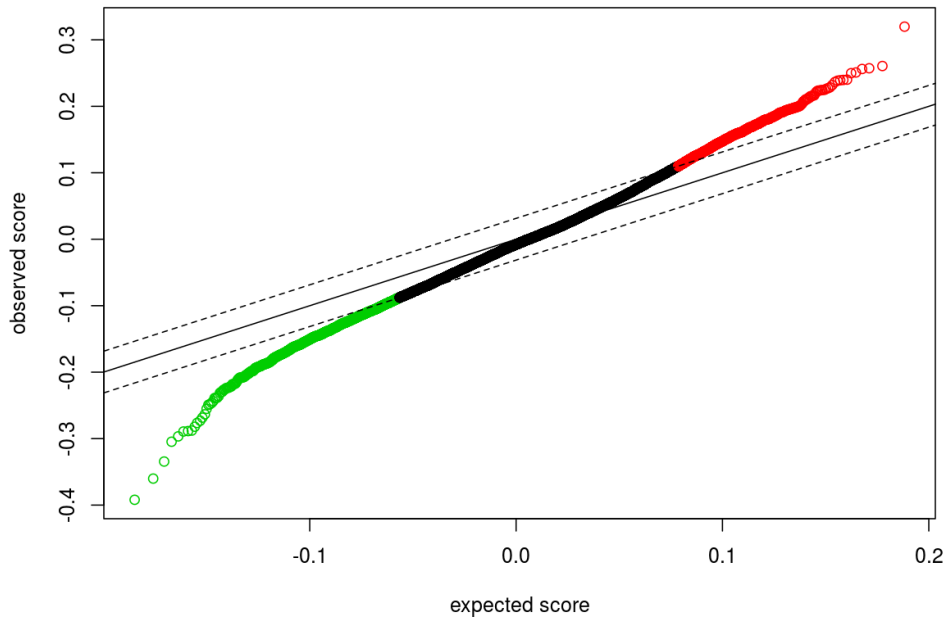


Illustration 17: Plot des gènes surexprimés (rouge) et sous-exprimés (vert) de l'objet samfit, créé par SAMseq. Cette analyse est réalisée avec l'âge de méthylation tumoral.

Accélération

UP RÉGULÉ		
Gene Symbol	Score(d)	**q-value(%)**
ENTPD1	0.341	**0**
NIPAL3	0.34	**0**
SH3BGRL2	0.335	**0**
APLP2	0.322	**0**
GXYLT2	0.313	**0**
RAD23B	0.311	**0**
SPRED2	0.311	**0**
C6orf174	0.31	**0**
C10orf72	0.306	**0**
FRMD3	0.305	**0**

DOWN RÉGULÉ		
Gene Symbol	Score(d)	**q-value(%)**
ADAMTS16	-0.315	**0**
VPREB3	-0.308	**0**
DDIT4L	-0.3	**0**
GSTK1	-0.297	**0**
TSPAN33	-0.294	**0**
MRPL16	-0.29	**0**
LIPC	-0.29	**0**
C1orf93	-0.284	**0**
CILP	-0.284	**0**
CCL28	-0.283	**0**

Tableau 19: Top 10 des gènes up-régulés et down-régulés pour l'accélération. Analyse réalisée avec SAMseq

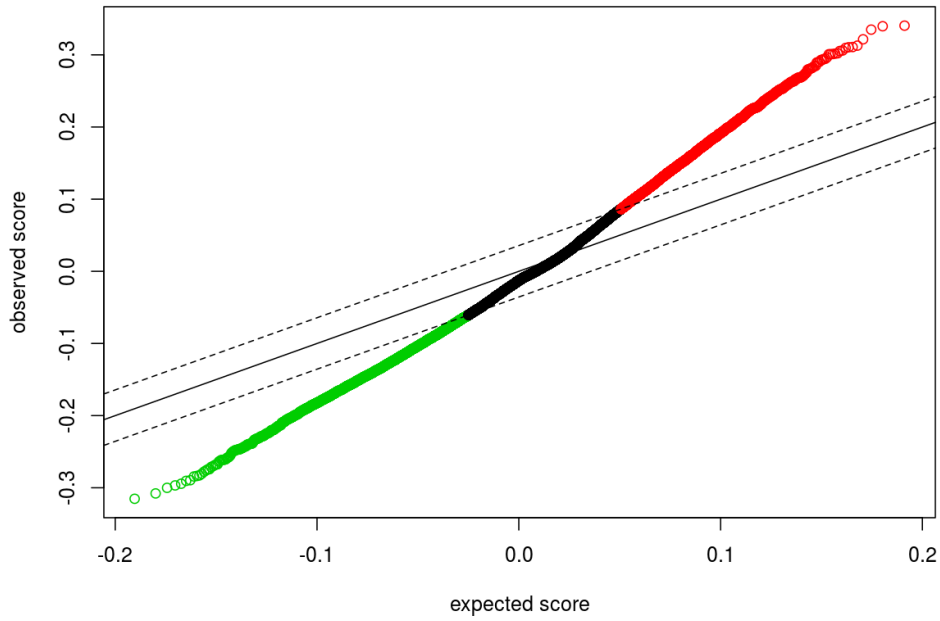


Illustration 18: Plot des gènes surexprimés (rouge) et sous exprimés (vert) de l'objet samfit, créé par SAMseq. Cette analyse est réalisée avec l'accélération.

Conclusion

SAMseq ressort de nombreux gènes légèrement surexprimés (en rouge sur le graphique) ou sous exprimés (en vert sur le graphique) par rapport à la variable analysée. Mais il est difficile de tirer des conclusions de ces analyses étant donné le grand nombre de gènes fournis par le programme. Cependant, nous ne remarquons aucun gène en commun dans les top 10 entre l'âge de méthylation tumoral et l'accélération.

Une analyse de ces résultats avec le logiciel DAVID permettrait de déterminer les pathways dont les gènes sont up ou down régulés.

CONCLUSION QUESTION 3

Les programmes GSEA et SAMseq demandent tous les 2 un formatage très précis des données. Il est donc important de leur fournir les fichiers dans le format adéquat, sans quoi aucune analyse n'est possible.

GSEA permet de déterminer les pathways pour lesquels une corrélation entre l'expression et la variable est observée. En effet, il se sert des « gene sets » de MsigDB. Cette database fournit pour un pathway les différents gènes connus pour y être impliqués. Dans notre cas, aucun pathway n'a pu être mis en évidence comme significativement corrélé avec une variable car le FDR reste toujours supérieur à 25 %. Nous avons donc trop de risques de faux positifs.

Quant à SAMseq, ce dernier fournit les gènes up ou down régulés en fonction de la variable analysée. De nombreux gènes semblent être modifiés mais leur analyse est complexe car ils sont très nombreux et nous ne savons pas les pathways dans lesquels ils sont impliqués. Ces résultats demandent donc une analyse supplémentaire avec un autre programme afin de déterminer les pathways dont les gènes sont différentiellement exprimés.

Nous avons donc pu nous initier à ces 2 programmes d'analyses de séquençage d'ARN et voir les avantages de chacun.

Bibliographie

Firehose Broad Institute : https://gdac.broadinstitute.org/

GSEA user guide : http://software.broadinstitute.org/gsea/doc/GSEAUserGuide.pdf

SAM used guide : https://cran.r-project.org/web/packages/samr/samr.pdf

Cours de V. Detours : BIOL-F423 Analysis of functional and comparative genomics data

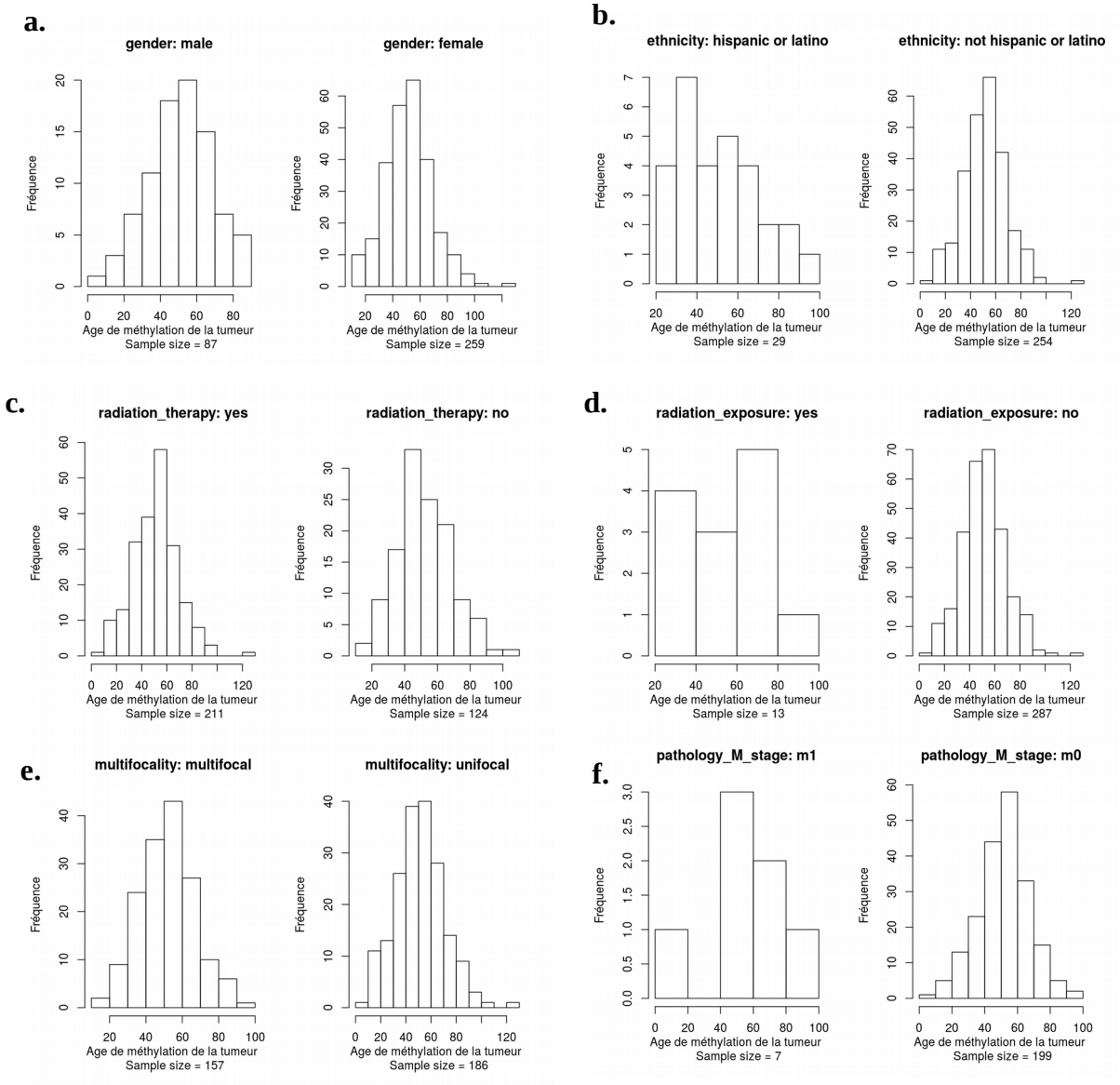
Programme R : R version 3.3.0 (2016-05-03)

Annexes

Question 2

Variables catégoriques en 2 groupes

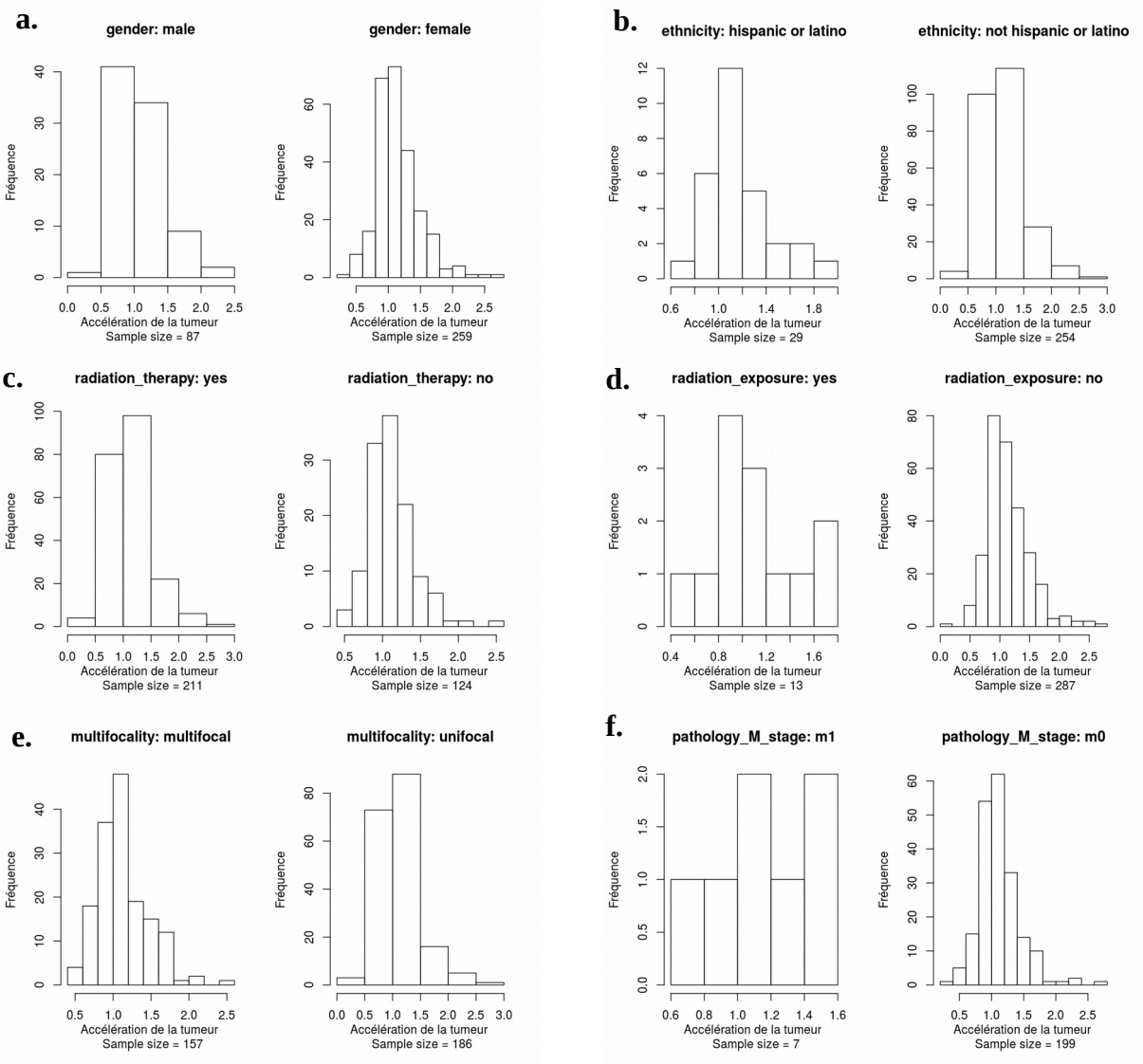
Age de méthylation tumoral, histogrammes



Annexe 1 : Histogrammes des paramètres cliniques catégoriques ayant 2 groupes pour l'analyse de l'ADN tumoral.

a. Genre, **b.** Ethnicité, **c.** Radiothérapie, **d.** Exposition aux radiations, **e.** Multifocalité et **f.** Stade pathologique M

Accélération, histogrammes



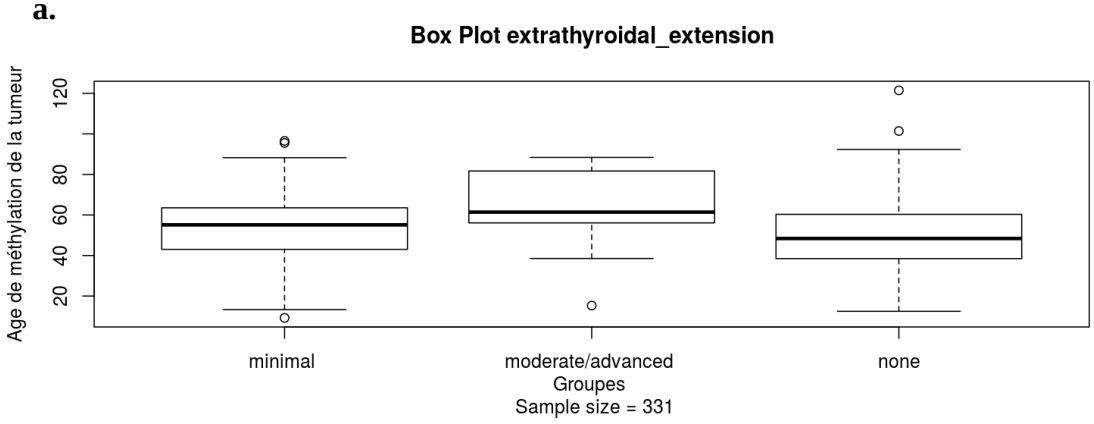
Annexe 2 : Histogrammes des paramètres cliniques catégoriques ayant 2 groupes pour l'analyse de l'ADN tumoral.

a. Genre, **b.** Ethnicité, **c.** Radiothérapie, **d.** Exposition aux radiations, **e.** Multifocalité et **f.** Stade pathologique M

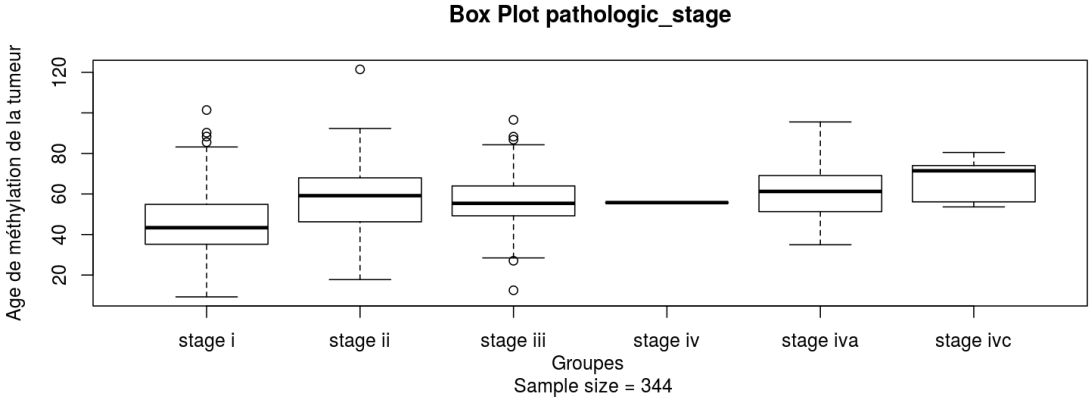
Variables catégoriques à plus de 2 groupes

Age de méthylation tumoral, box plots

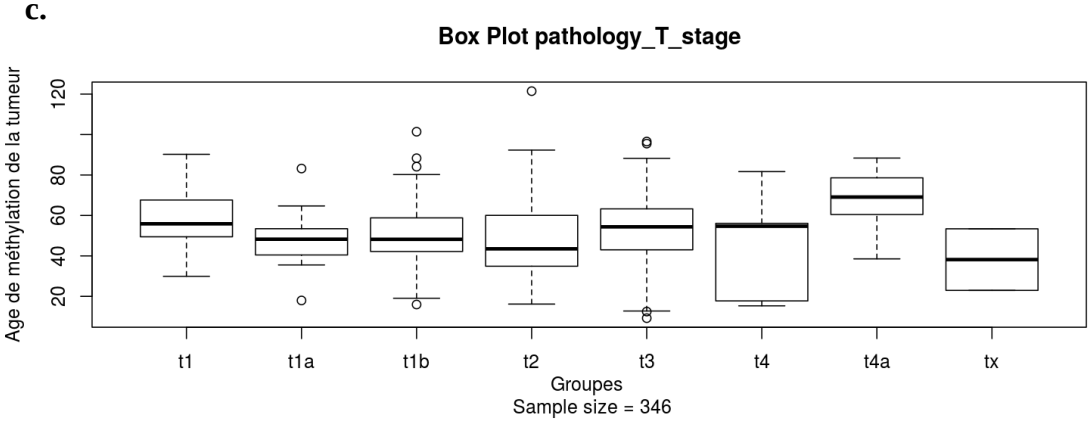
a.



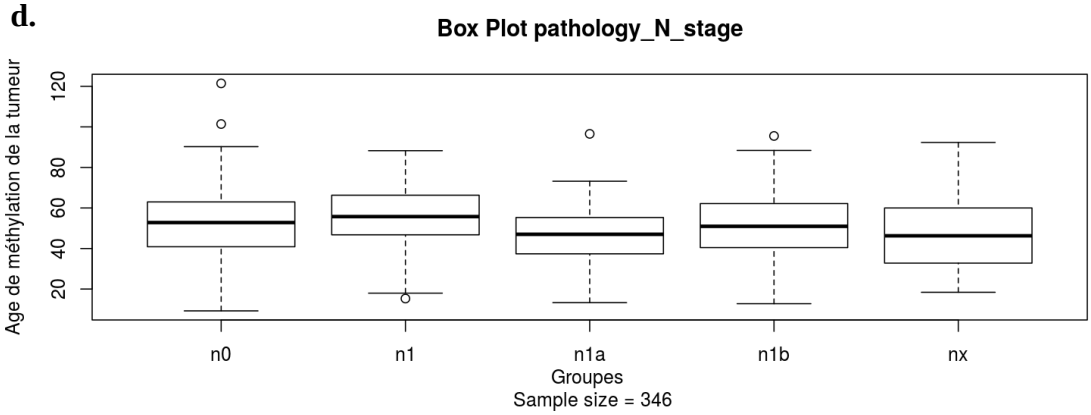
b.



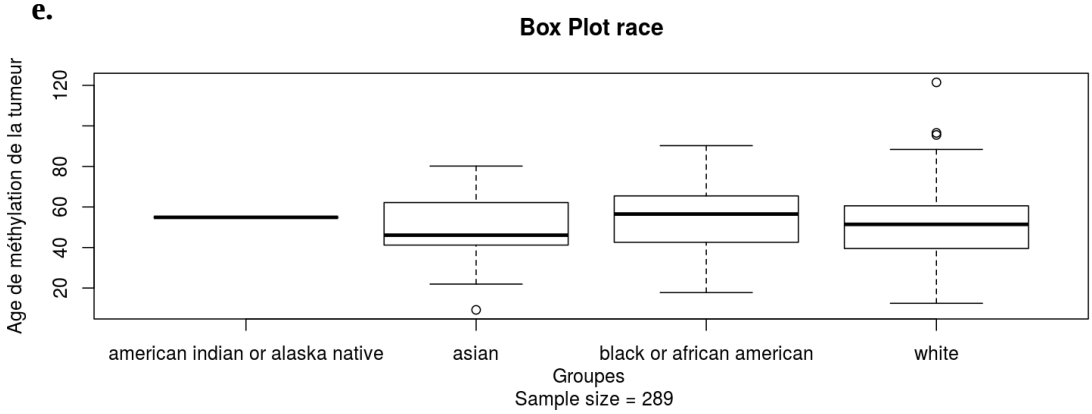
c.



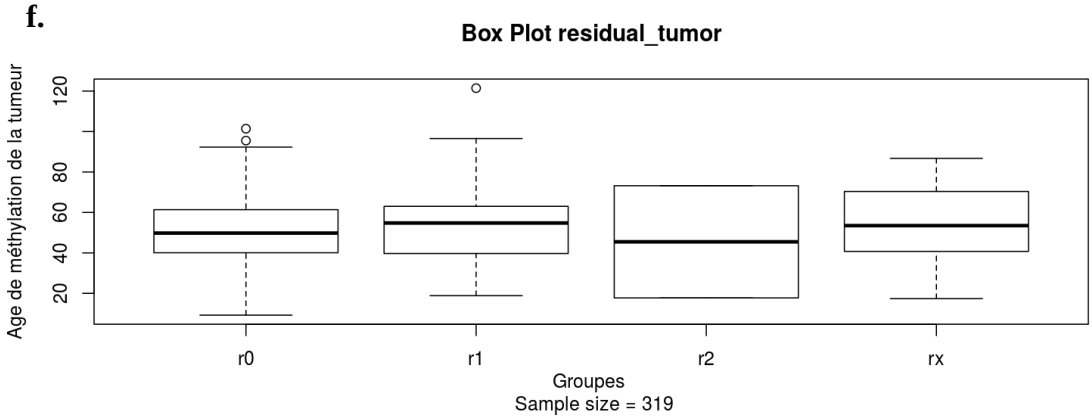
d.



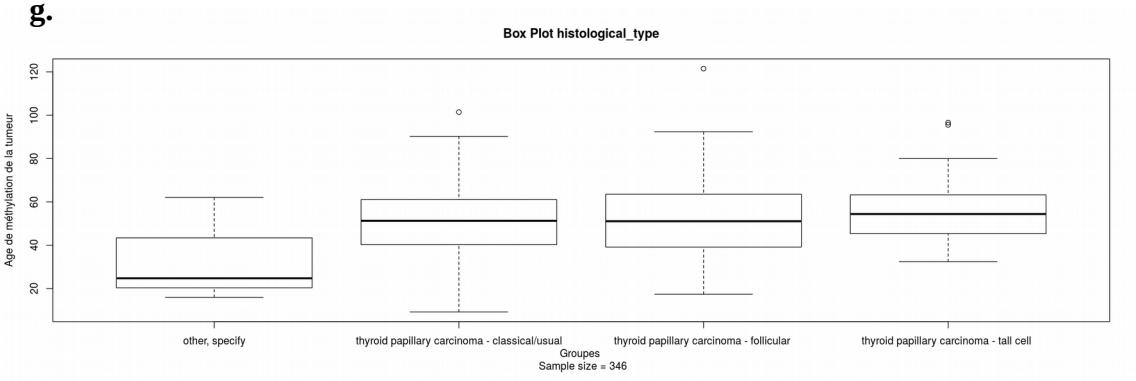
e.



f.



g.

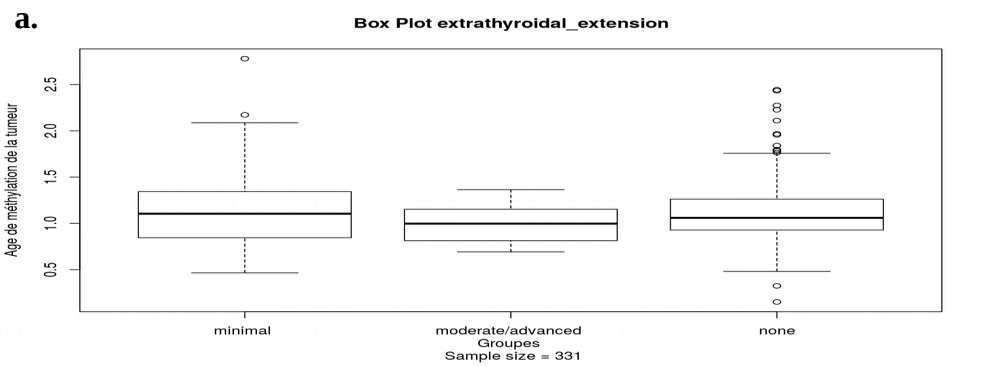


Annexe 3 : Box plots des paramètres cliniques catégoriques ayant plus de 2 groupes pour l'analyse de l'ADN tumoral.

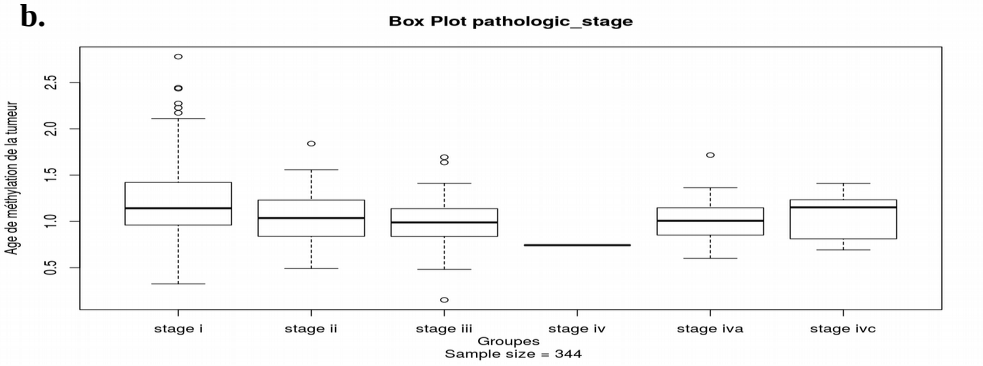
a. Extension tumorale, **b.** Stade pathologique, **c.** Stade pathologique T, **d.** Stade pathologique N, **e.** Race, **f.** Tumeur résiduelle et **g.** Type histologique

Accélération, box plots

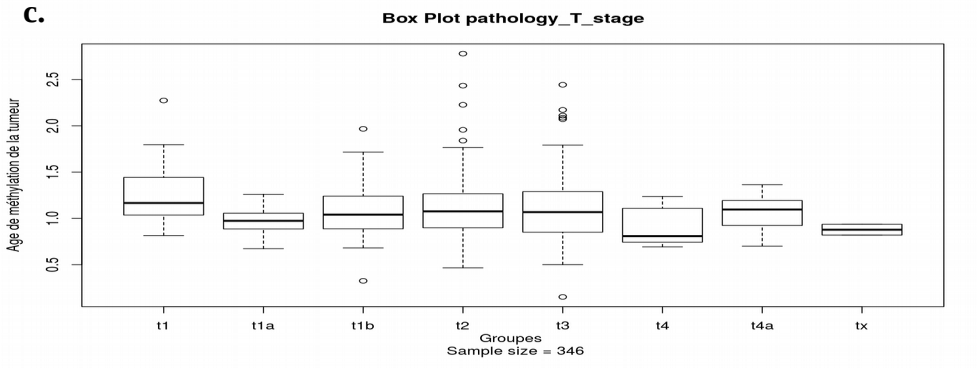
a.



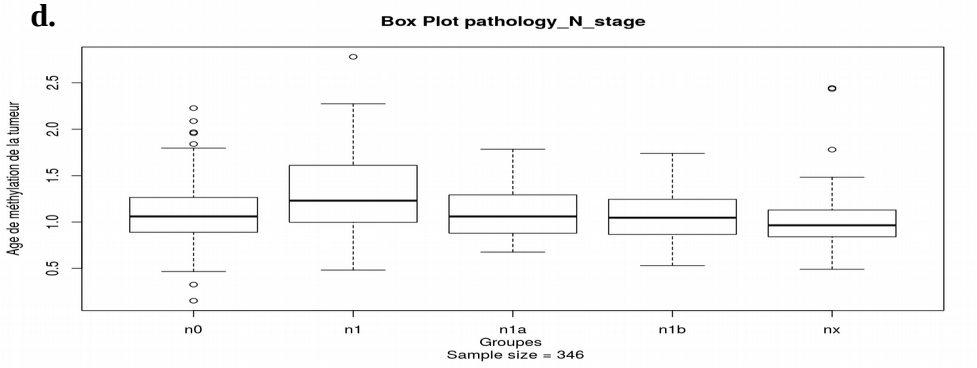
b.



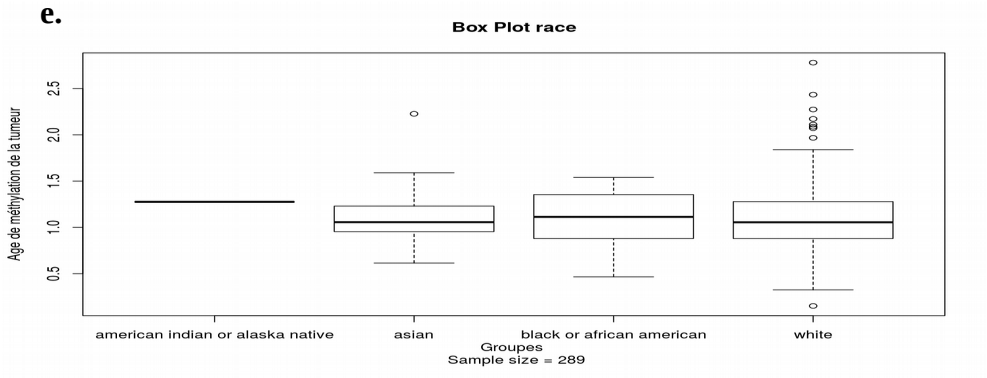
c.



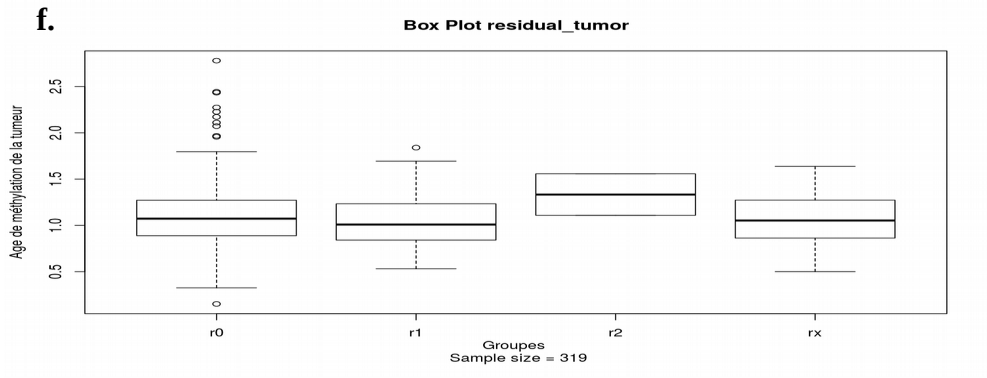
d.



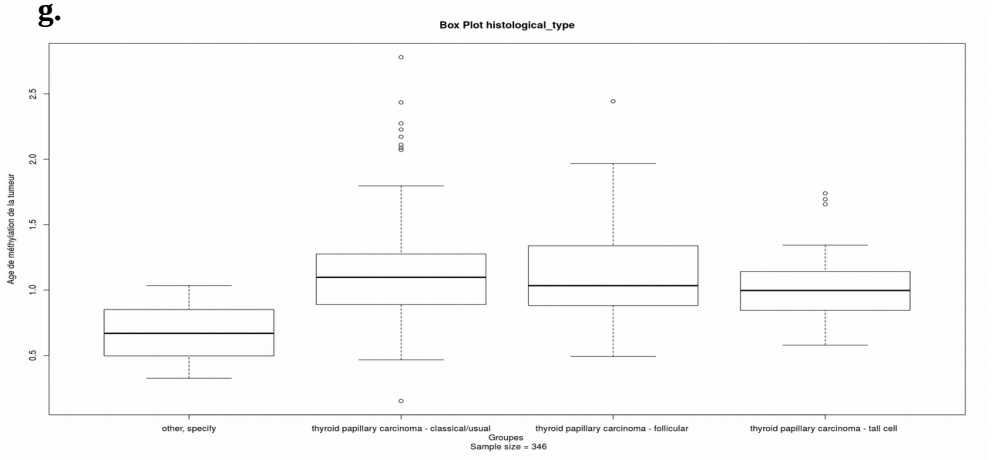
e.



f.



g.



Annexee 4 : Box plots des paramètres cliniques catégoriques ayant plus de 2 groupes pour l'analyse de l'accélération.

a. Extension tumorale, **b.** Stade pathologique, **c.** Stade pathologique T, **d.** Stade pathologique N, **e.** Race, **f.** Tumeur résiduelle et **g.** Type histologique