

BIOL-F-423 : Project 2016

Vincent Detours - projet

Introduction

Pour réaliser ce travail, j'ai travaillé sur les données COAD - Colon adenocarcinoma - extraites de la base de données de Broad GDAC firehose.

L'adénocarcinome du côlon est un cancer assez fréquent chez les personnes âgées entre 40 et 60 ans (1). Nous allons analyser l'âge de méthylation dans les cellules cancéreuses et comparer cet âge de méthylation avec l'âge chronologique des patients. Les données ont été analysées avec le programme R. Le pseudo-code qui m'a permis de réaliser ce travail a été fourni dans les annexes.

Question 1:

Afin de réaliser ces tests, nous avons dû extraire les données d'âge des patients sains et les données d'âge des patients atteints du cancer. L'ensemble des données a été extrait avec le programme R. Pour l'ensemble des scatter-plots ci-dessous, la corrélation a été mesurée ainsi que la significativité de celle-ci (i.e. calcul de p-value). Le programme R nous permet d'effectuer ces tests statistiques facilement grâce à des fonctions spécifiques (e.g. `cor.test(x,y, method=c("spearman"))`).

Lien entre âge chronologique et âge de méthylation de l'ADN

Afin de tester, la corrélation qu'il existe entre l'âge de méthylation et l'âge chronologique, nous avons utilisé le test de Spearman qui mesure la dépendance statistique entre 2 valeurs.

Dans nos données, nous avons 87 données de cellules saines et 343 cellules tumorales.

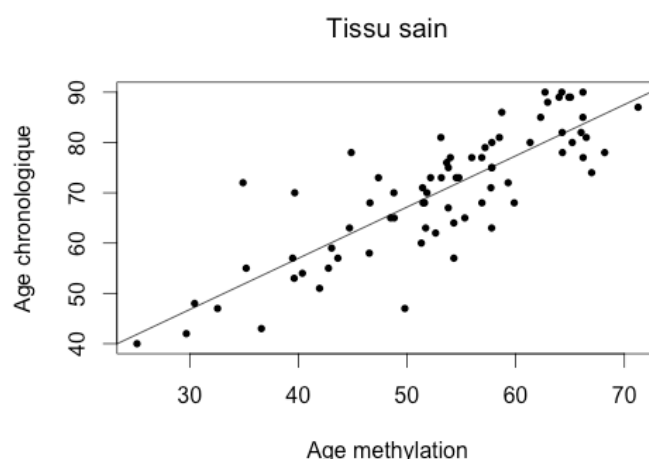


Figure 1: comparaison de l'âge de méthylation des tissus sains et de l'âge chronologique

Nous avons comparé l'âge de méthylation et l'âge chronologique dans les cellules des tissus sains des patients. Nous avons obtenu un coefficient de corrélation de 0,7351

et une p-value de $5.191e-16$. Nous pouvons dire qu'il y a une corrélation entre ces 2 paramètres. On peut ajouter qu'utiliser l'âge de méthylation des tissus sains permet une estimation tout aussi correcte de l'âge des patients par rapport à l'âge chronologique.

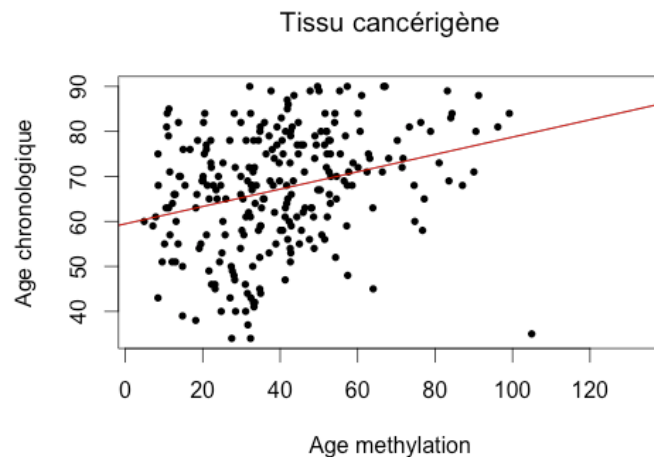


Figure 2: comparaison de l'âge de méthylation des tissus tumoraux et de l'âge chronologique

En ce qui concerne les tissus cancérigènes, la courbe des tendances montre que les cellules ont tendance à avoir un âge différent de l'âge réel des patients. Il semblerait que les cellules cancéreuses soient plus "jeunes".

Cependant, lorsqu'on observe le coefficient de corrélation entre ces 2 paramètres, il est de $0.2466 > 0,05$. La corrélation n'est pas significative. De plus, R nous indique qu'il n'est pas possible de calculer exactement la p-value exacte à cause d'ex-aequo.

Sur la figure ci-dessous, nous tentons de déterminer si l'âge de méthylation des tissus tumoraux est corrélé à l'âge de méthylation des tissus sains pour les mêmes patients.

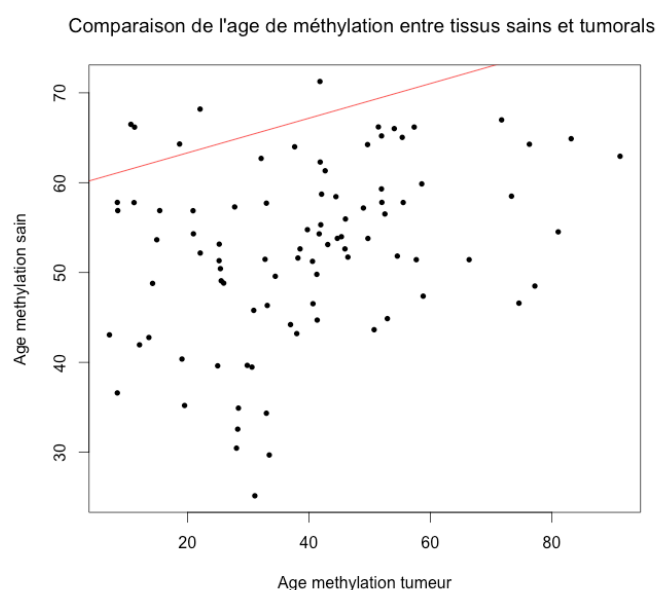


Figure 3: comparaison entre l'âge de méthylation des tissus tumoraux et de l'âge de méthylation des tissus sains par patients

On observe que la corrélation est plus faible que précédemment (corrélation = 0.52, p-value=6,213 e-06). Les cellules tumorales auraient tendance à avoir leur ADN plus sujette à la méthylation que les cellules saines et seraient donc plus "vieilles" chez les patients atteints d'un cancer.

Accélération

Ensuite, nous allons analyser 2 manières de calculer l'accélération de l'âge tumoral. L'accélération peut être définie soit comme le rapport de DNAm age tumor/chronological age ou comme le rapport DNAm age tumor/ DNAm age normal.

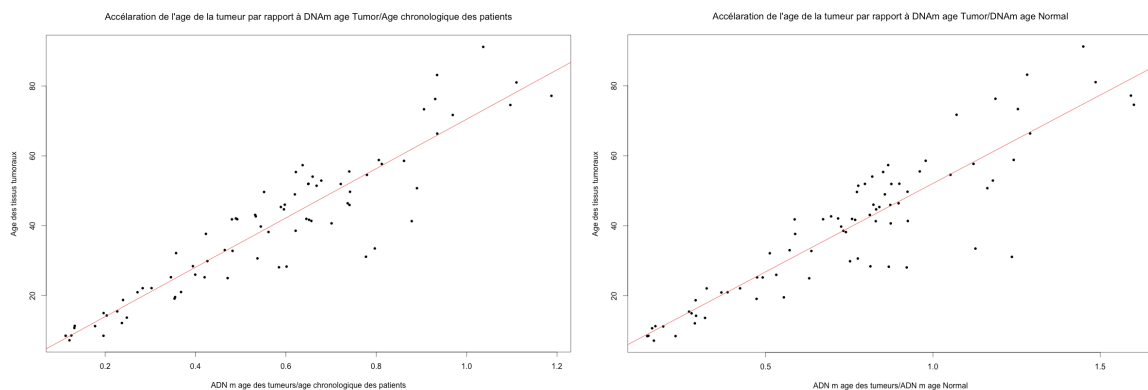


Figure 4: comparaison de l'âge des tumeurs par rapport à DNAm age Tumor/Âge chronologique des patients (corrélation = 0.90 et p-value < 2.2e-16), comparaison de l'âge des tumeurs par rapport à DNAm age Tumor/ DNAm age Tumor (corrélation = 0.87 et p-value < 2.2e-16)

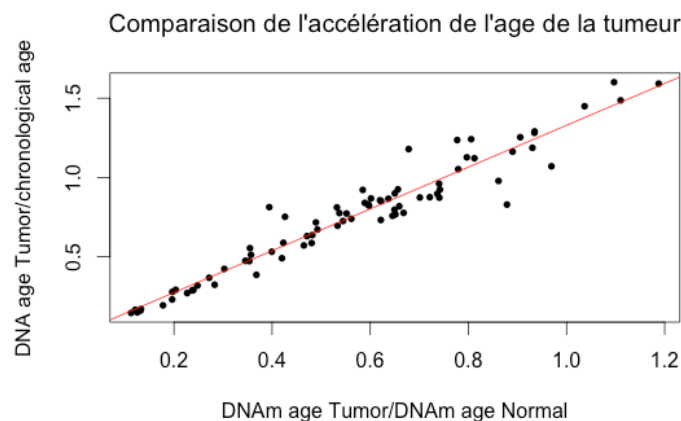


Figure 5: comparaison entre les 2 différentes manières de calculer l'accélération de l'âge de l'ADN

La corrélation est de 0.96 entre les 2 rapports avec une p-value de 2.2e-16 (fig5). La corrélation étant élevée et significative, nous aurions pu prendre n'importe quelle manière de calculer l'accélération. Cependant, lorsqu'on regarde la figure 4, on observe

que la corrélation entre le rapport DNAm age Tumor/Age chronologique et l'âge des tumeurs est plus élevé. Il est donc préférable d'utiliser ce rapport pour calculer l'accélération pour nos données. Nous utiliserons donc ce rapport pour analyser les données cliniques (voir Q2).

On peut retirer des graphes ci-dessus que l'âge de méthylation des tissus sains correspond relativement bien à l'âge chronologique des patients. Tandis que l'âge de méthylation des tissus tumoraux correspond moins à l'âge chronologique des patients malade. Le cancer pourrait donc avoir un effet de sur la méthylation de l'ADN et donc sur l'âge biologique des patients.

Q2 :

Pour cette question, nous avons comparé l'âge de méthylation de l'ADN des cellules tumorales en fonction des différentes variables cliniques.

J'ai sélectionné les variables qui me semblaient les plus pertinentes. La variable concernant le site de la tumeur n'a pas été analysée car toutes les mesures se trouvaient au niveau du colon. De plus, d'autres variables quant à elles comptaient de nombreux NA (e.g. days_to_last_knows_alive).

Ces comparaisons vont permettre de vérifier s'il y a des différences d'âge de méthylation des cellules tumorales en fonction des classes des variables cliniques. En fonction du nombre de moyenne à comparer, nous utiliserons soit les tests de student et de Wilcoxon-Mann-Witney (pour analyser 2 moyennes), soit l'ANOVA et le test de Kruskal-Wallis (pour comparer plus de 2 moyennes) :

-H₀: les moyennes sont identiques

-H₁: les moyennes ne sont pas identiques

On rejettera l'hypothèse nulle si la p-value < 0,05.

Les premiers box-plots ont été réalisés en comparant l'âge de méthylation des cellules tumorales aux variables cliniques, la section suivante sera consacrée à l'accélération.

Comparaison en fonction des stades

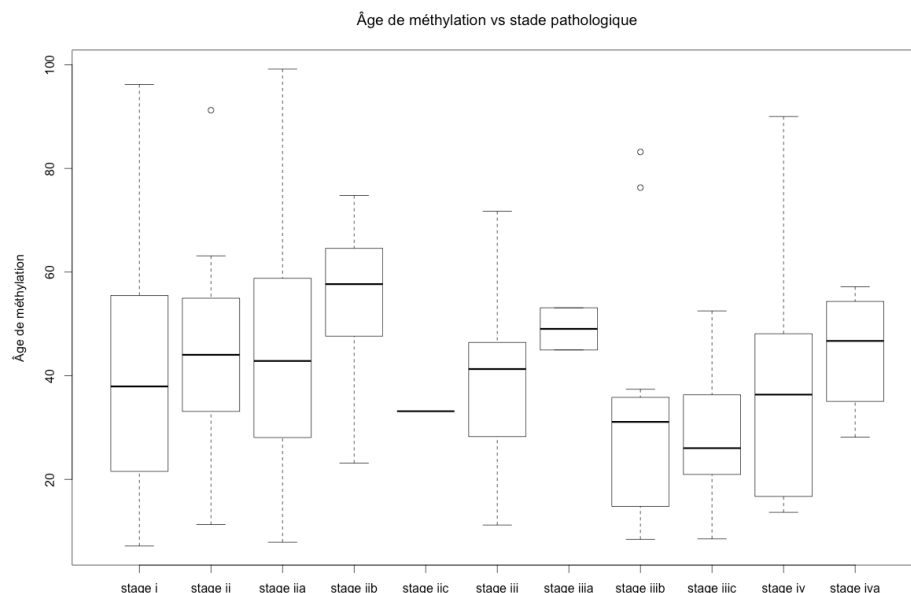


Figure 6 : comparaison de l'âge de méthylation et du stade pathologique (Pr(>F) =0.247)

Sur la figure 6, nous avons une distribution en fonction des stades pathologiques. Après avoir effectué le test de l'ANOVA, on peut voir que nous avons une Pr value supérieure à 0,05. Il n'y a pas de corrélation entre l'âge de méthylation et le stade pathologique.

Dans nos données, nous avons aussi une précision au sein des stades pathologiques (en fonction de l'état T, de l'état N et de l'état M).

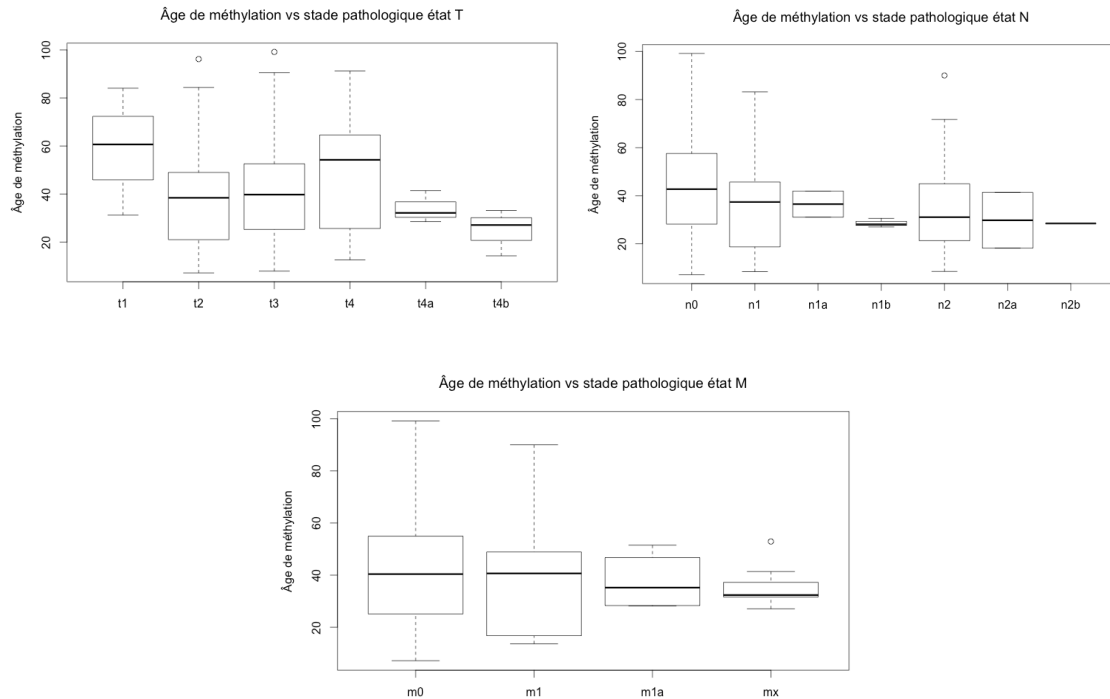


Figure 7: comparaison de l'âge de méthylation et du stade pathologique état T ; ($\text{Pr}(> F) = 0.261$), comparaison de l'âge de méthylation et du stade pathologique état N ; ($\text{Pr}(> F) = 0.263$), comparaison de l'âge de méthylation et du stade pathologique état M ; ($\text{Pr}(> F) = 0.76$),

Pour les boxplots des différents stades, nous observons une Pr value supérieure à 0.05, il n'y a donc pas de corrélation entre l'âge de méthylation des cellules tumorales et leurs différents stades.

Nous aurions peut-être pu prendre plus de données pour obtenir des graphes significatifs.

Comparaison en fonction d'autres variables

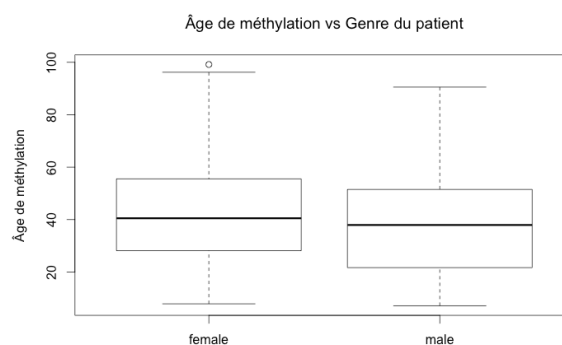


Figure 8: comparaison de l'âge de méthylation et le genre ; ($\text{Pr}(> F) = 0.15$)

Sur la figure 8, nous constatons qu'il n'y a pas de relations entre le genre des patients et l'âge de méthylation des tissus tumoraux.



Figure 9: comparaison de l'âge de méthylation et le type de tissu histologique ; ($\text{Pr}(>F) = 0.00188$)

En ce qui concerne le type de tissus histologiques, on peut observer qu'il y a une différence significative en fonction du type de tissus. Les tissus du colon ont un âge de méthylation moins élevé que les tissus de types colon mucineux. Une recherche plus poussée quant aux fonctions de ces deux types cellulaires pourrait être réalisée afin de trouver une raison à cette différence d'âge de méthylation.

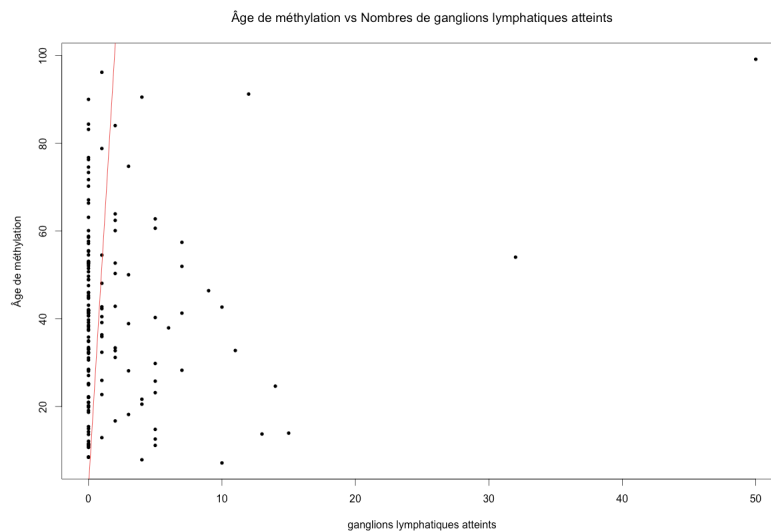


Figure 9: comparaison de l'âge de méthylation et le nombre de ganglions lymphatiques atteints ; $p\text{-value} = 0.939$ et $\text{corrélation} = -0,0062$

On a comparé le nombre de ganglions lymphatiques atteints et l'âge de méthylation. Vu que nous avons 2 valeurs quantitatives, il était préférable de réaliser un scatter plot. Vu que la $p\text{-value} > 0.05$, la corrélation n'est pas significative. Il n'y a pas de relation entre le nombre de ganglions atteints et l'âge de méthylation.

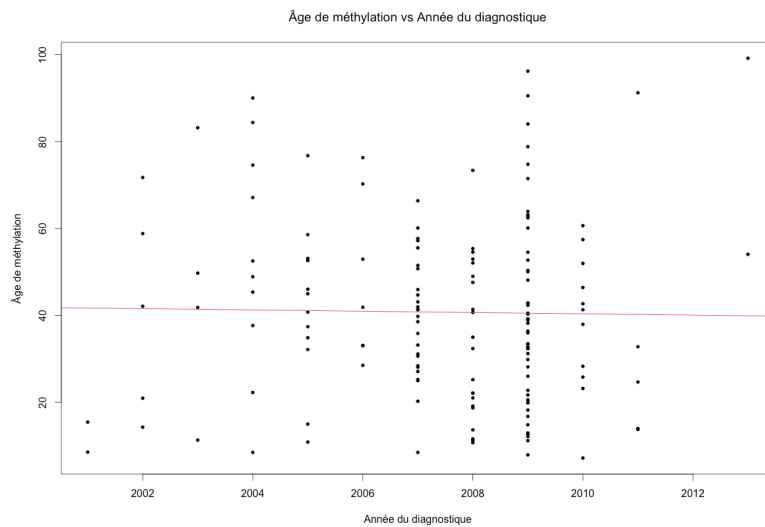


Figure 10: comparaison de l'âge de méthylation et l'année du diagnostique ; $p\text{-value} = 0.835$ et $\text{corrélation} = -0,00618$

Comme précédemment, il était préférable d'utiliser un scatter plot. Ici encore, il n'y a pas de corrélation entre les 2 variables.

Avec l'accélération

Pour cette partie ci, nous avons comparé les données précédentes au rapport DNAm age Tumor/age chronologique (voir Question 1).

Comparaison en fonction des stades

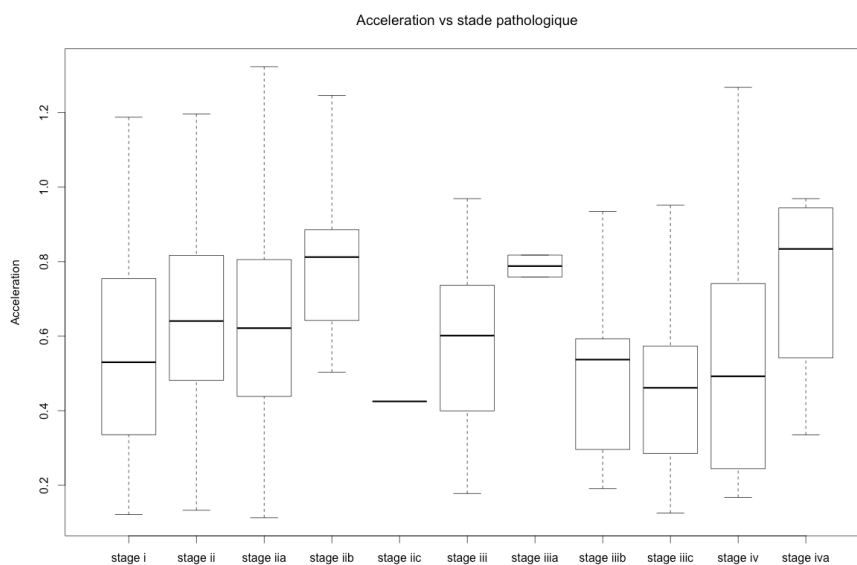


Figure 11: comparaison de l'accélération et du stade pathologique ($\text{Pr}(> F) = 0.264$)

Comme pour la figure 6, il ne semble pas avoir de corrélation entre les stades pathologiques et accélération.

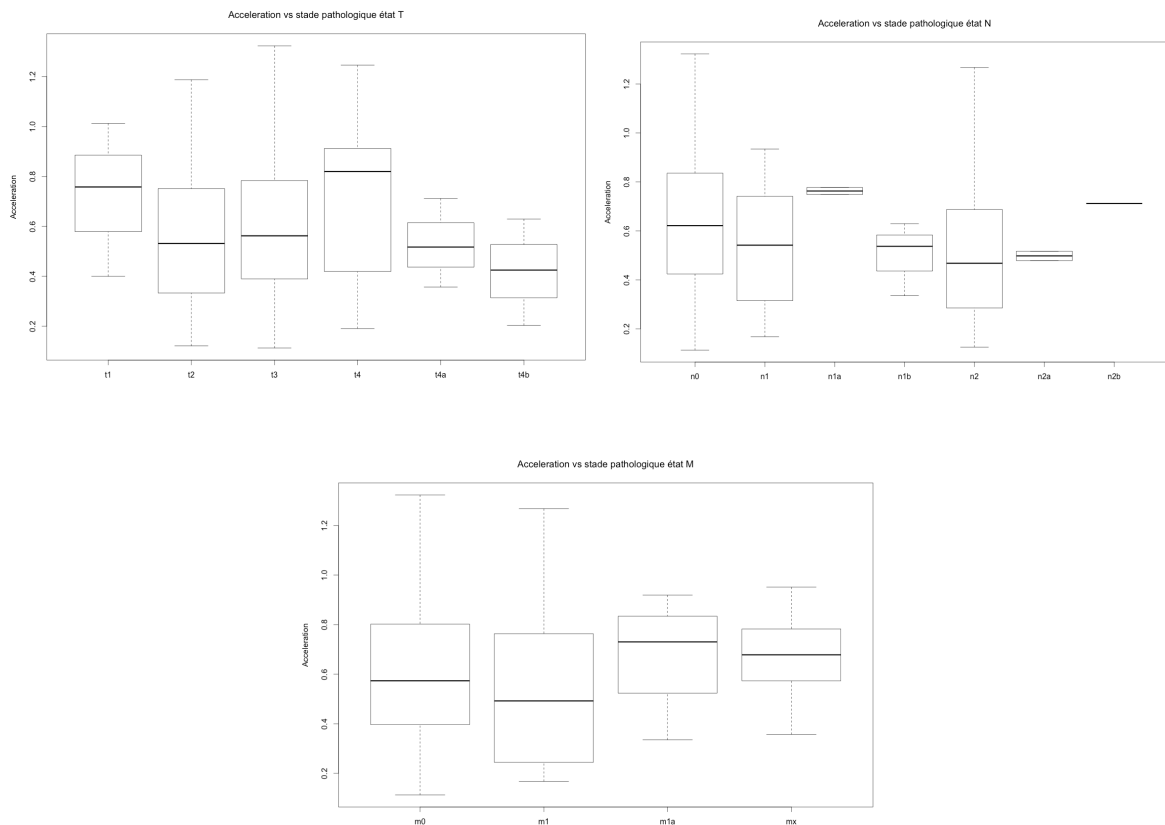


Figure 12: comparaison de l'accélération et du stade pathologique état T ; ($\Pr(>F) = 0.385$), comparaison de l'accélération et du stade pathologique état N ; ($\Pr(>F) = 0.76$), comparaison de l'accélération et du stade pathologique état M ; ($\Pr(>F) = 0.119$),

En prenant l'accélération pour comparer nos données, nous avons toujours de P-value supérieure à 0,05. Beaucoup de variables étaient sans valeurs, il est donc possible que le manque de significativité provienne de ce manque de valeur.

De plus, en sachant qu'on avait une corrélation 0.9 entre l'accélération et l'âge de méthylation absolue des cellules tumorale, on pouvait s'attendre à observer des résultats similaires.

Comparaison en fonction d'autres variables

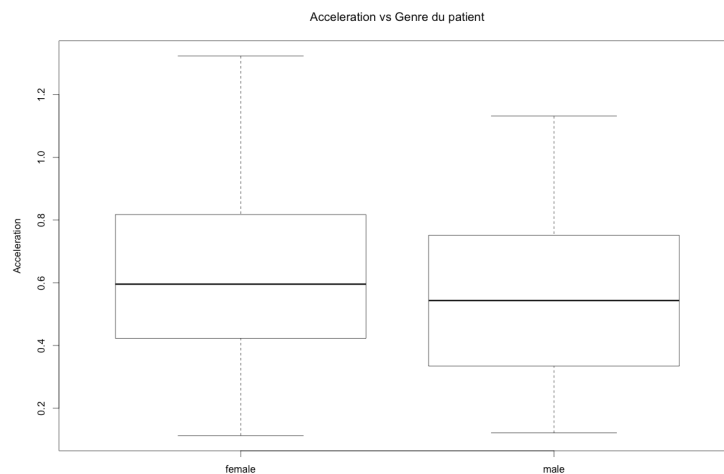


Figure 13: comparaison de l'accélération et le genre ; ($\Pr(>F) = 0.119$)

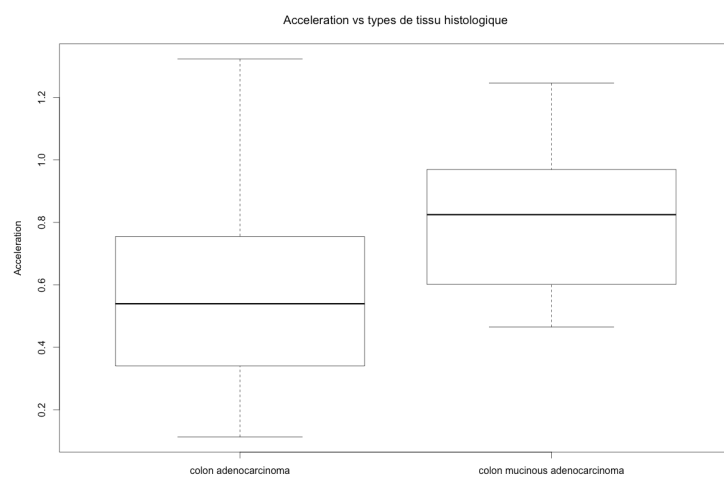


Figure 14: comparaison de l'accélération et le type de tissu histologique ; ($\Pr(>F) = 0.000433$)

Comme précédemment, uniquement la comparaison avec le type de tissu histologique est significative.

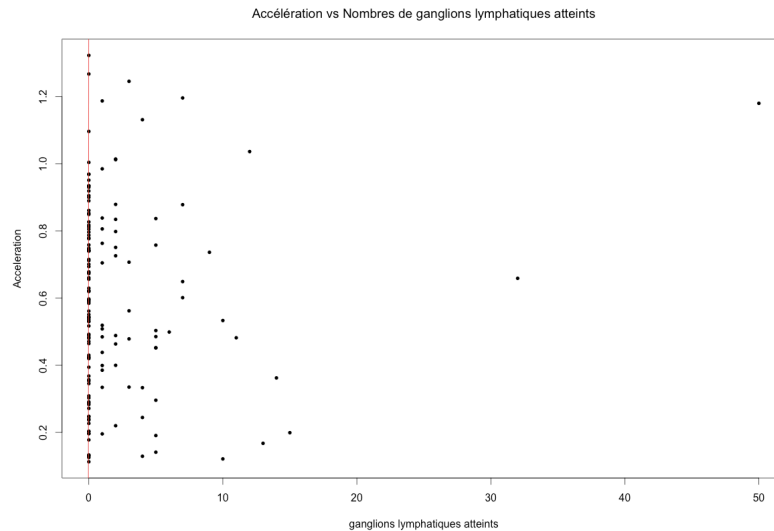


Figure 15: comparaison de l'accélération et le nombre de ganglions lymphatiques atteints ; p-value = 0.329 et corrélation = -0,0019

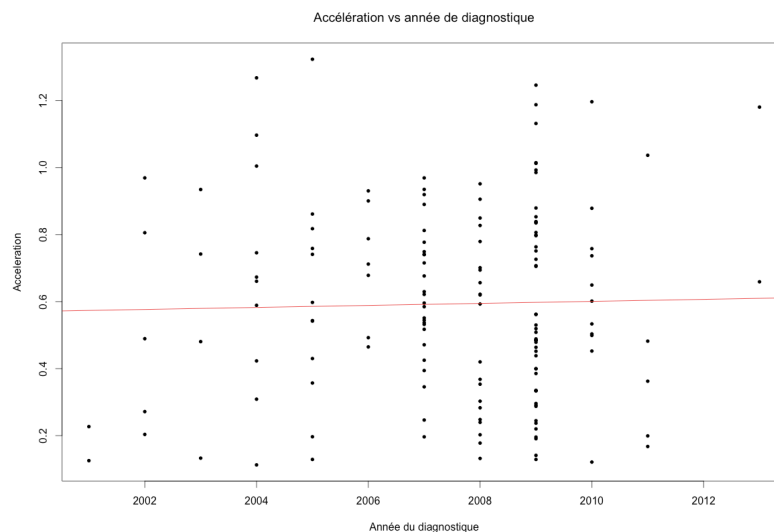


Figure 16: comparaison de l'accélération et l'année de diagnostique du cancer ; p-value = 0.75 et corrélation = -0,0019

Nous avons des résultats similaires aux comparaisons avec l'âge de méthylation absolue.

Après une analyse des graphiques, il semblerait que l'utilisation de l'accélération amplifie les différences entre les graphiques. Même si nous n'avons pas pu observer des changements significatifs entre les 2 mesures, les différences entre les graphiques semblent plus marquées. Une étude plus poussée aurait été nécessaire pour conforter cette intuition (e.g. avec plus de données de patients différents,...).

Analyse de survie

Dans cette partie, nous allons étudier les courbes de survie. Afin de mesurer la significativité des différentes courbes, nous avons utilisé les tests logrank. Une analyse de survie permet de modéliser le temps qu'il reste avant la mort d'un organisme biologique.

Le test du logrank permet de savoir si les 2 courbes sont superposées. Lorsque le test est significatif, H_0 est rejetée.

H_0 : Les courbes sont superposées

H_1 : Au moins une 1 courbes n'est pas superposées

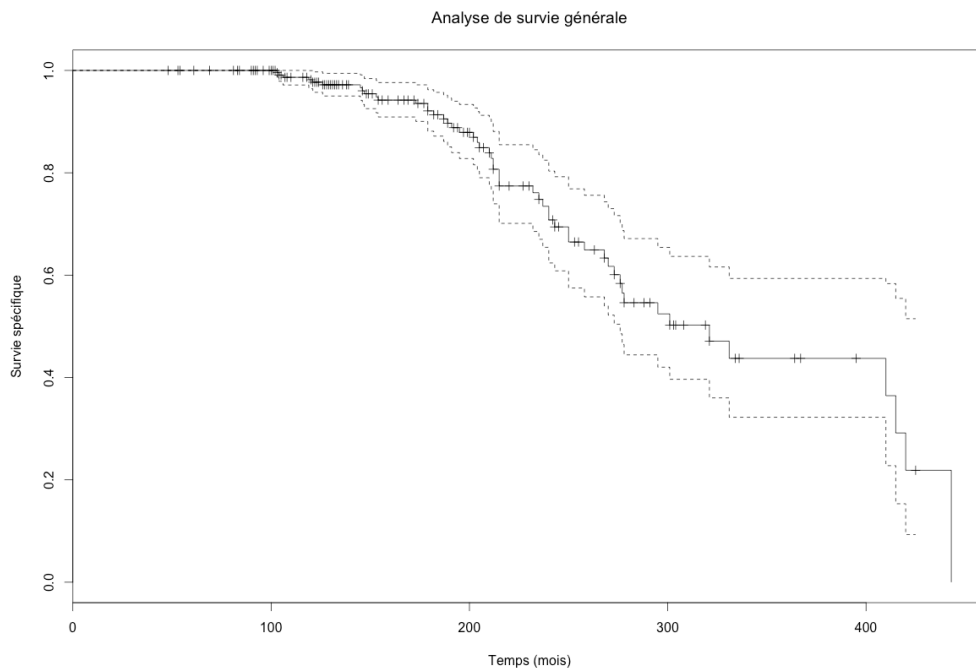


Figure 17: Analyse de survie des patients atteints d'un cancer du côlon

Les croix se trouvant le long de la courbe de survie représentent des censures. Une censure apparaît lorsqu'on arrête de suivre un patient avant qu'il soit décédé.

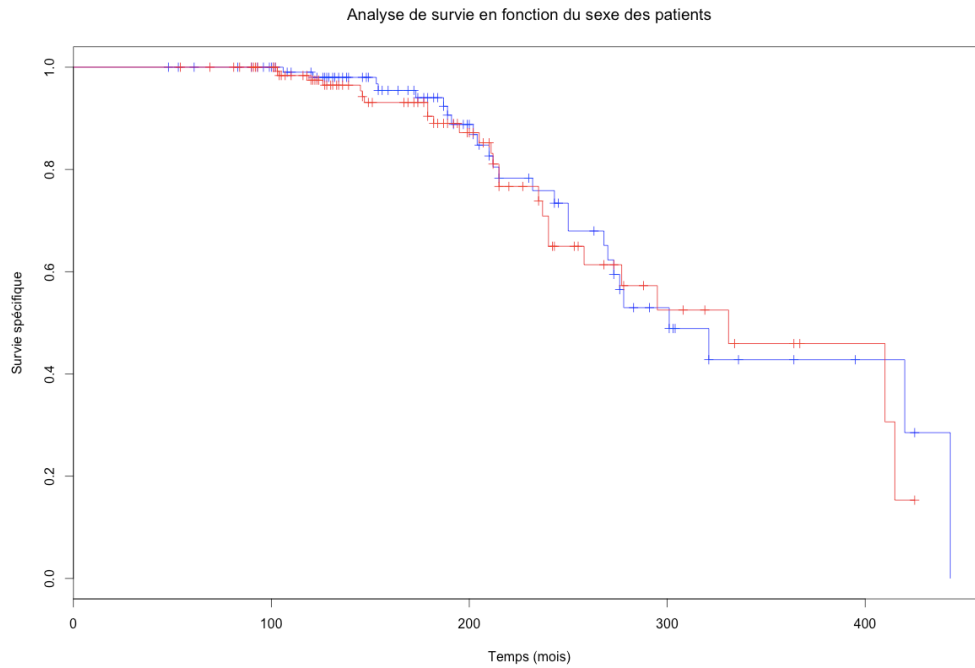


Figure 18: Analyse de survie des patients atteint d'un cancer du côlon en fonction du sexe.
Bleu : homme et Rouge : femme

Le test logrank de la figure 18 nous donne une $p\text{-value} = 0.7094$. Nous pouvons pas rejeter l'hypothèse nulle. Il n'y a pas de différence de survie entre les sexes.

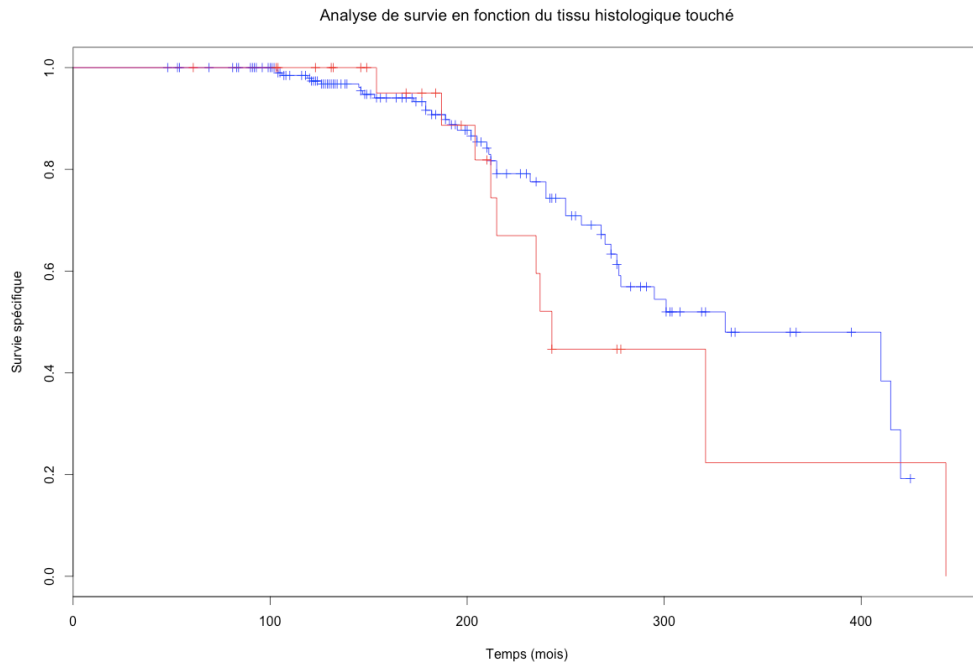


Figure 19: Analyse de survie des patients atteint d'un cancer du côlon en fonction du type du tissu histologique touché.
Bleu : colon adénocarcinoma et Rouge : colon mucinous adénocarcinoma

Sur la figure 19, la $p\text{-value} = 0,3045$, il n'y a pas de différence significative dans la survie du patient en fonction du types de tissus histologiques touchés.

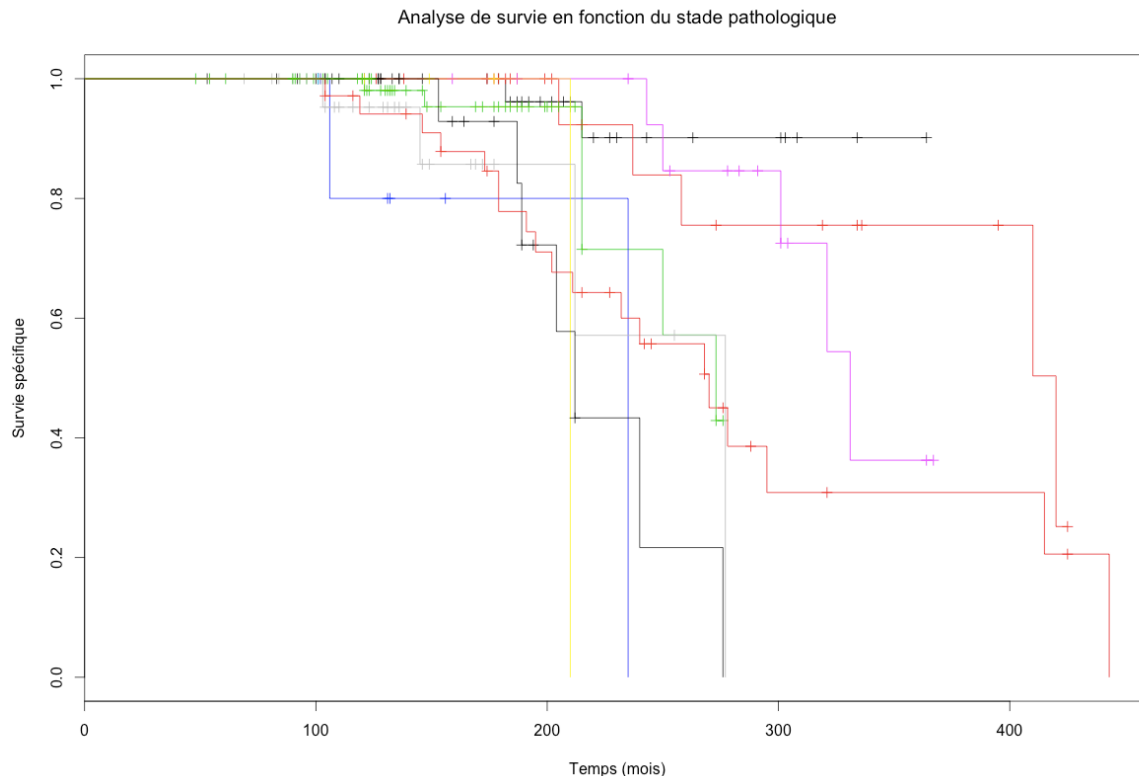


Figure 20: Analyse de survie des patients atteint d'un cancer du côlon en fonction du stade pathologique.

Couleur non spécifiée - voir tableau ci-dessous

	coef	exp(coef)	se(coef)	z	Pr(> z)
survivalTable[, 9]stage ii	6.531e-01	1.921e+00	8.438e-01	0.774	0.438943
survivalTable[, 9]stage iia	1.504e+00	4.501e+00	8.233e-01	1.827	0.067706 .
survivalTable[, 9]stage iib	3.193e+00	2.437e+01	1.011e+00	3.160	0.001578 **
survivalTable[, 9]stage iic	NA	NA	0.000e+00	NA	NA
survivalTable[, 9]stage iii	8.184e-01	2.267e+00	8.402e-01	0.974	0.330025
survivalTable[, 9]stage iiia	2.847e+00	1.723e+01	1.240e+00	2.295	0.021730 *
survivalTable[, 9]stage iiib	2.396e+00	1.098e+01	8.725e-01	2.746	0.006033 **
survivalTable[, 9]stage iiic	2.713e+00	1.508e+01	8.095e-01	3.352	0.000803 ***
survivalTable[, 9]stage iv	1.841e+00	6.300e+00	7.508e-01	2.451	0.014229 *
survivalTable[, 9]stage iva	-1.246e+01	3.881e-06	4.759e+03	-0.003	0.997911

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Comme on peut le voir dans le tableau ci-dessus, certains stades sont significatifs par rapport à d'autre. Les chances de survie seront différentes si le patients se trouvent aux stades iib , iiia, iiib, iiic, iv.

Analyse de survie en fonction des dernières nouvelles du patient (last follow up dans le tableau). On peut voir que plus le temps est long avant le dernier suivis plus le patient a des chances de décéder.

Il aurait été intéressant d'effectuer les tests de survie en fonction de la radiothérapie mais la grande majorité des patients n'en suivaient pas et ou les données étaient manquante.

Q3

Pour la question 3, le programme GSEA a d'abord été téléchargé. Ensuite, les données ont été triées pour correspondre au format attendu par GSEA.

GSEA va nous permettre de savoir si des gènes sont exprimés significativement par rapport à d'autres.

Avant de pouvoir utiliser GSEA, nous avons traité nos données pour que celles-ci correspondent au format attendu (voir http://www.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats).

Coller L-Tracer Fusionner													
A1 geneD													
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	geneD	TCGA.A6.267	TCGA.A6.268	TCGA.A6.268	TCGA.A6.268	TCGA.A6.268	TCGA.A6.268	TCGA.A6.268	TCGA.A6.268	TCGA.A6.268	TCGA.A6.268	TCGA.A6.268	TCGA.A6.268
2	A1BG	41.00	54.00	83.88	70.00	155.79	18.00	67.02	40.00	65.00	108.96	54.46	29.38
3	A1CF	283.00	490.00	155.00	278.00	306.00	431.00	62.00	607.00	249.00	387.00	690.00	307.00
4	A2BP1	19.00	0.00	1.00	61.00	0.00	3.00	0.00	5.00	159.00	19.00	2.00	0.00
5	A2LD1	235.19	344.90	265.80	272.52	190.03	129.23	277.16	372.47	551.80	575.07	1089.17	515.26
6	A2ML1	0.00	0.00	0.00	0.00	19.00	0.00	0.00	2.00	1.00	1.00	1.00	1.00
7	A2M	21170.95	7499.94	17165.87	33768.91	12598.86	13121.95	6453.99	8084.97	7877.96	13509.94	10244.95	8238.93
8	A4GALT	572.00	432.00	502.00	881.00	943.00	470.00	205.00	217.00	50.00	316.00	201.00	193.00
9	A4GNT	0.00	1.00	2.00	2.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
10	AAA1	3.00	4.00	6.00	2.00	31.00	0.00	2.00	0.00	0.00	0.00	4.00	0.00
11	AAAS	1414.00	1930.00	1982.00	1403.00	1448.00	2482.00	1965.00	1758.00	1610.00	3081.00	2010.00	1933.00
12	AACSL	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	2.00
13	AACS	1639.00	2376.00	2465.00	1522.00	2310.00	2067.00	1666.00	2047.00	2020.00	2817.00	2239.00	2997.00
14	AADACL2	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
15	AADACL3	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
16	AADACL4	0.00	2.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
17	AADAC	3.00	127.00	0.00	1465.00	164.00	13.00	52.00	7.00	8.00	49.00	29.00	12.00
18	AADAT	302.00	338.00	160.00	182.00	85.00	383.00	419.00	317.00	349.00	517.00	275.00	616.00

Fichier experience.coad formaté pour GSEA

	A	B	C	D	E	F	G	
1	#numeric							
2	#tumorDNAm							
3	41.8479123752643	39.7450552587744	34.9049269393589	28.3841273120356	47.3683606702619	53.799963005002	30.4509773548503	53.11
4	#acceleration							
5	10.4619780938161	9.93626381469361	6.98098538787177	7.0960318280089	15.789453556754	17.9333210016673	7.61274433871256	10.62
6								
7								
8								

Fichier phenotypelabel formaté pour GSEA

Malgré un formatage précis des données, un chargement des données réussi sur GSEA et de nombreux tests (voir annexe). Je n'ai pas réussi à obtenir une analyse par GSEA.

SAM

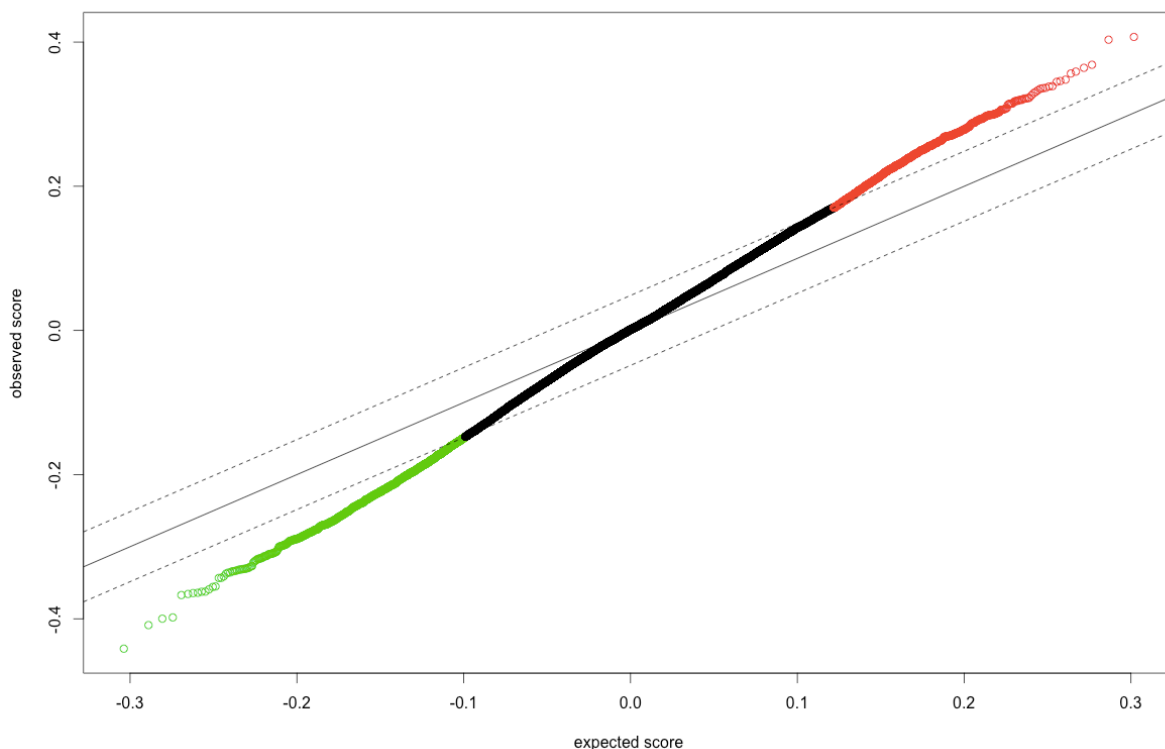


Figure 21 : graphique SAM créé par SAMseq avec les données d'accélération.

Le graphique SAM permet de visualiser les gènes surexprimés (rouges) et les gènes sous-exprimés (vert). Cependant, vu le nombre élevé de gènes, il est difficile de tirer des conclusions spécifiques au cancer du côlon.

TOP 20 gènes les plus exprimés				TOP 20 des gènes les moins exprimés			
Gene ID	Gene Name	Score(d)	q-value(%)	Gene ID	Gene Name	Score(d)	q-value(%)
[1,]	"ZFYVE28"	"57732"	"0.428" "0"	[1,]	"NHLRC1"	"378884"	"-0.418" "0"
[2,]	"MBP"	"4155"	"0.426" "0"	[2,]	"SLC22A11"	"55867"	"-0.393" "0"
[3,]	"MAPRE2"	"10982"	"0.421" "0"	[3,]	"GPR143"	"4935"	"-0.388" "0"
[4,]	"GAREM"	"64762"	"0.41" "0"	[4,]	"NUBPL"	"80224"	"-0.384" "0"
[5,]	"KCNK6"	"9424"	"0.406" "0"	[5,]	"SLC39A2"	"29986"	"-0.381" "0"
[6,]	"MAP3K6"	"9064"	"0.404" "0"	[6,]	"CELP"	"1057"	"-0.379" "0"
[7,]	"SMAD2"	"4087"	"0.403" "0"	[7,]	"ZNF3"	"7551"	"-0.378" "0"
[8,]	"TFF2"	"7032"	"0.394" "0"	[8,]	"GYG2"	"8908"	"-0.374" "0"
[9,]	"RNF125"	"54941"	"0.393" "0"	[9,]	"DCAF4"	"26094"	"-0.374" "0"
[10,]	"SGMS2"	"166929"	"0.392" "0"	[10,]	"SLC25A26"	"115286"	"-0.372" "0"
[11,]	"TBC1D8"	"11138"	"0.391" "0"	[11,]	"AK3"	"50808"	"-0.371" "0"
[12,]	"SMCHD1"	"23347"	"0.382" "0"	[12,]	"TFAM"	"7019"	"-0.371" "0"
[13,]	"ASPHD2"	"57168"	"0.38" "0"	[13,]	"AXIN2"	"8313"	"-0.369" "0"
[14,]	"RALBP1"	"10928"	"0.38" "0"	[14,]	"APCDD1"	"147495"	"-0.368" "0"
[15,]	"CCDC109B"	"55013"	"0.379" "0"	[15,]	"CALCA"	"796"	"-0.367" "0"
[16,]	"C18orf25"	"147339"	"0.379" "0"	[16,]	"LFNG"	"3955"	"-0.365" "0"
[17,]	"GIPR"	"2696"	"0.379" "0"	[17,]	"GBAS"	"2631"	"-0.364" "0"
[18,]	"SLC9A1"	"6548"	"0.376" "0"	[18,]	"REPIN1"	"29803"	"-0.364" "0"
[19,]	"TCP11L2"	"255394"	"0.374" "0"	[19,]	"RP9"	"6100"	"-0.363" "0"
[20,]	"KDELR3"	"11015"	"0.372" "0"	[20,]	"RPS6KA6"	"27330"	"-0.362" "0"

Sur le tableau ci-dessus, nous avons répertorié les 20 gènes les plus exprimés et les 20 gènes les moins exprimés. Si nous avions obtenu les résultats de GSEA nous aurions pu

les comparer à ceux-ci. De plus, nous aurions pu analyser le "rôle" de ces gènes, leurs fonctions au niveau des tissus du côlon ou dans les cancers de manière générale.

Bibliographie

1. https://fr.wikipedia.org/wiki/Cancer_du_côlon consulté le 1 juin 2016
2. Jerome Torrisani and Frederic Lopez. Methylation de l'adn et regulation epigenetique des cancers. Hepato-Gastro et Oncologie Digestive, 2003.
3. Ahuja N and Issa JP. Aging, methylation and cancer. Histol Histopathol.

Annexes

```
#Pseudo-code

#extraction des données

matrice_donnee_age_methyl <- load("COADREAD-5.Rda")
""" extractions des données d'age de méthylation """
matrice_donnee_variable_clinique <- load("COAD.Clinical_Pick_Tier1")
""" extraction des variables cliniques des patients """

#Parsing des données en fonction des tissus sains ou tumoraux
""" Si code bar compris entre 1 et 9, la donnée est issue d'un tissu tumoral. Si code bar compris entre 10 et 19, la
donnée est issue d'un tissu sain. """

healthy_table=[[bar_code],[tissu sain],[age_methylation],[age_chronologique]]
tumor_table=[[bar_code],[tissu tumoral],[age_methylation],[age_chronologique]]

#####
#Q1
#####

for i in range(len(matrice_donnee_age_methyl)):
    if 1 <= matrice_donnee_age_methyl[bar_code] <= 9 and
    matrice_donnee_age_methyl[bar_code] == matrice_donnee_age_methyl[bar_code] :
        tumor_table[age_methyl]= matrice_donnee_age_methyl[age_methyl]
    if 10 <= matrice_donnee_age_methyl[bar_code] <= 19 and
    matrice_donnee_age_methyl[bar_code] == matrice_donnee_age_methyl[bar_code] :
        healthy_table[age_methyl]= matrice_donnee_age_methyl[age_methyl]

plot(x= age_methylation,y=age_chronologique) #graphique pour chacune des tables
cor.test(x,y) #calcul de la corrélation

#rajout des colonnes pour acceleration

for each patient with DNA_age_methyl_tumor and DNA_age_methyl_healthy :
    table creation with [bar_code,age_chrono, DNA_age_methyl_tumor, DNA_age_methyl_healthy,
    DNA_age_methyl_tumor/age_chrono, DNA_age_methyl_tumor/DNA_age_methyl_healthy]

plot(x=acceleration, y= DNA_age_methyl_tumor) # graphique pour chacune des accélération
cor.test(x,y)
plot(x=acceleration1,y=acceleration2) #graphique entre les 2 types d'accélération
cor.test(x,y)

#####
#Q2
#####

for clinical_variable in matrice_donnee_variable_clinique :
```

```

boxplot(DNA_age_methyl_tumor, clinical_variable)
anova(DNA_age_methyl_tumor~clinical_variable)

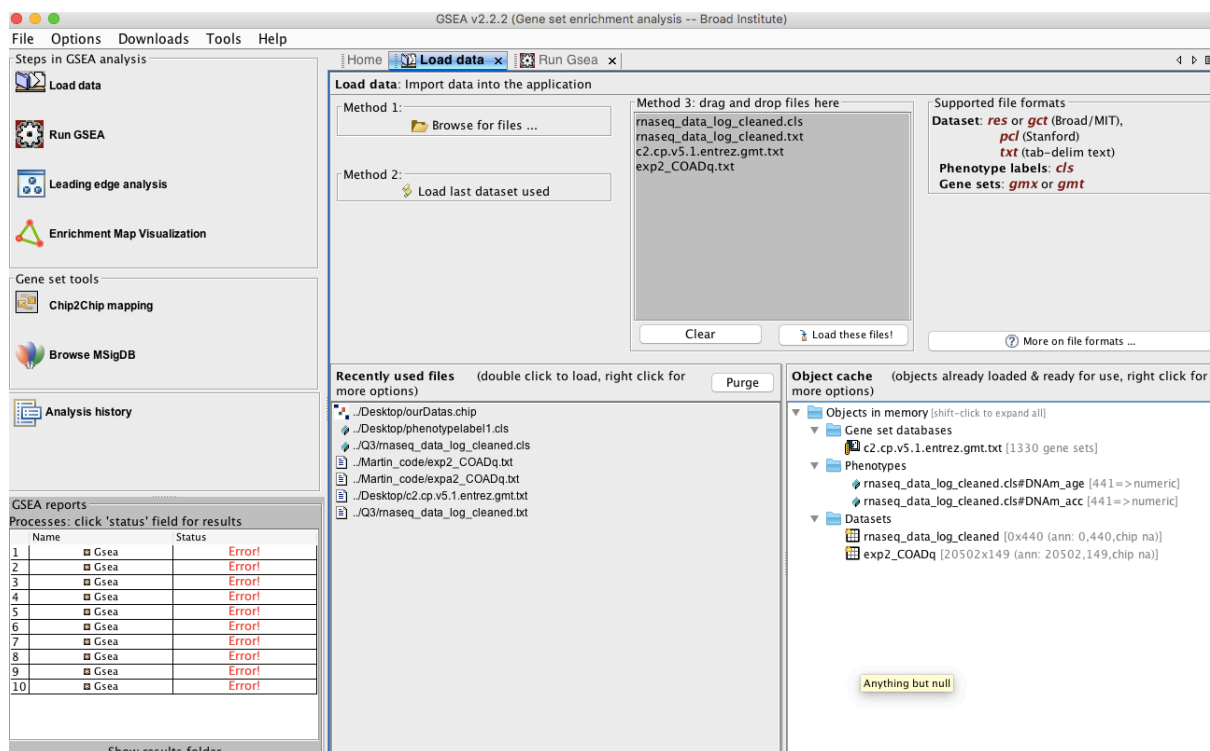
for clinical_variable in matrice_donnee_variable_clinique :
    boxplot(acceleration, clinical_variable)
    anova(acceleration~clinical_variable)

#survival plot

for each patient in matrice_donnee_variable_clinique:
    survival_table cration with [patient_id, initial_diagnosis, days_to_init_diag,days_from_last_news,
    months_from_last_news,vital_status, gender, Tissus, stade pathologique]

s=surv(months_from_last_news, vital_status)
plot(surfit(s~1))
plot(surfit(s~gender))
coxph(s~gender)
plot(surfit(s~Tissus))
coxph(s~tissus)
plot(surfit(s~ stade pathologique))
coxph(s~stade pathologique)

```



Annexe 2 : programme GSEA avec données chargées