DNA methylation and gene expression in Colorrectal Adenocarcinoma

Olga Ibáñez Solé

Abstract
The relationships between the DNA methylation age computed as described by Horvath (2013), the DNAm age acceleration and several clinical variables were explored in colorectal adenocarcinoma samples. Two different measurements of the DNAm acceleration (with respect to the chronological age of the patients and with respect to the DNAm age of patient-matched healthy tissues) were considered. The DNA methylation age appeared to be severely disrupted in colorrectal adenocarcinoma tissues, and slightly decelerated in colorrectal healthy tissues. Survival analyses showed that there was no difference in survival depending on the patients being subject to radiation therapy, the size and the hystological type of the tumor or the sex of the patient. There was no significant difference in survival between the half of the patients whose tumor tissues had the strongest DNAm acceleration and the half with the lowest DNAm acceleration. The Gene Set Enrichment Analysis showed that very few gene sets were significantly correlated with DNAm age and DNAm age acceleration. The gene sets that were significantly correlated with the DNAm acceleration depended on the expression of the acceleration used. Many of the gene sets for which a significant correlation was found in terms of nominal p-value were not considered significant at the FDR $<$ 25% level.

Keywords
Colorrectal carcinoma — DNA methylation clock

Introduction

DNA methylation in Colorrectal Adenocarcinoma

Colorrectal cancer is one of the leading causes of cancer-related deaths worldwide, and is related to well characterized DNA methylation profiles (Fernandes et al, 2017). Colon and rectum cancers develop from premalignant lesions in the epithelium that turn into malignant adenocarcinomas and might ultimately lead to metastasis.

Colorrectal tumors, like other solid tumors, are characterized by a widespread hypomethylation, with hypomethylated blocks that are up to hundreds of Kb long. This feature is thought to be related to the tumorigenesis onset and progression through the promotion of genomic instabillity and the activation of abnormal genes (Fernandes et al, 2017).

Aberrant hypermethylation of specific CpG islands has also been observed in colorrectal cancers. A particular phenotype named the CpG island methylator phenotype (CIMP) is described as a distinguishable subset of colorrectal cancers that have a characteristic hypermethylation of five genes (CACNA1G, IGF2, NEUROG1, RUNX3 and SOCS1). CIMP-positive and CIMP-negative tumors have been shown to have fundamental clinicopathological differences (Fernandes et al, 2017).

Overall, DNA methylation appears to be one of the most interesting and relevant epigenetic marks in colorrectal cancers, and much effort is being put into unraveling the links between specific methylation patterns and the clinical characteristics associated to them.

DNA methylation age

Horvath proposed in 2013 a multi-tissue age predictor based in the DNA methylation degree of several genes. He developed this predictor using 8000 samples from 82 Illumina DNA methylation array datasets, which included 51 healthy tissues and cell types. He used a pensalized regression on a transformed version of the chronological age of the subjects and the CpGs. The model selected 353 CpGs which are at the basis of the DNAm clock. Horvath studied the predicted DNA methylation age (hereafter: DNAm age) on a large set of healthy and cancer tissues, and concluded that the DNA methylation age is extremely well corelated with the chronological age in healthy tissues and that it is usually accelerated in cancers, though it can be decelerated in some tissues and cancers as well (Horvath, 2013).

DNA methylation age acceleration

In the present work we used the predicted DNAm age of tumor and healthy tissues and studied the correlation between those and the chronological age of the patients. We also defined two different expressions for the DNAm age acceleration: 1) as the acceleration with respect to the DNAm age of healthy samples, and 2) with respect to the chronological age.

$$acceleration = \frac{DNAm\ age(tumor)}{DNAm\ age(healthy)} \quad (1)$$

$$acceleration = \frac{DNAm\ age(tumor)}{Chronological\ age} \quad (2)$$

We will hereafter refer to these two definitions of the DNAm age acceleration as expressions (1) and (2).

We studied the link between the DNAm age acceleration and several clinical variables: whether the patient had undergone radiation therapy, the size of the main tumor, its hystological type, the site where that tumor was located (either in the colon or in the rectum), and the sex of the patient.

We performed survival analyses on the patients according to the different categorical variables. We also converted the continuous variables DNAm age, DNAm acceleration (1) and DNAm acceleration (2) into categorical variables by splitting the patients in two halfs (low and high) according to the values of each of those three variables, and compared the survival probabilities of each of the groups.

We then performed a GSEA using the DNAm age and the two different expressions of the DNAm age acceleration as continuous covariates.

Gene Set Enrichment Analysis

GSEA is a bioinformatics tool designed to analyse genome-wide expression data on a gene-set level. That is to say, instead of searching for individual genes that are differentially expressed, it searches for the differential expression of ensembles of genes sharing a common function or belonging to the same pathway. It thus tries to overcome some of the limitations of studying the expression of individual genes, and there are a number of reasons to support this approach (Subramanian et al, 2005):

First of all, the fact that biologically relevant changes in in the expression of individual genes may be overlooked as they might not be statistically significant after the pertinent significance corrections for multiple hypothesis testing have been applied.

Second, the fact that giving a biological interpretation to the upregulation of a few unrelated genes is difficult, and scientists might be tempted to focus on those genes that are the most related to their area of expertise.

Third, a 20-fold overexpression of an individual gene might not be as biologically relevant as a more modest tuning in the regulation of a set of genes encoding for members of the same metabolic pathway. That is, cellular processes often involve the regulation of ensembles of genes rather than individual genes.

Last but not least, it has been shown that different studies on the same biological system performed by separate groups show little overlap in the resulting significantly regulated genes.

In the present work we perform a GSEA using the genome-wide expression data of 37 tumor samples and using the DNAm age and the DNAm age acceleration (1) and (2) as continuous covariates, to see whether there are specific groups of genes that are differentially expressed in concordance with their DNAm age or acceleration.

1. Methods

Data preprocessing

The DNAm ages of the patients computed according to the DNAm clock described by Horvath (2013) were provided by Prof. Detours. The clinical annotations of the patients as well as the mRNA expressions estimated from TCGA mRNA-seq data were downloaded from the Firehose website. The raw dataset contained measurements of 15 clinical variables from 352 colorrectal adenocarcinoma and 87 healthy tissue samples.

Missing and duplicated data treatment

The two samples for which the chronological age was missing were deleted from the dataset (IDs: ”AATA” and ”6889”). The rest of the missing values in the dataset (NAs in clinical variables) were not considered in the calculations.

There were three patients for which we had two different tumor tissue DNAm age measurements (IDs: ”2677”, ”3809” and ”3810”). In these three cases the duplicated samples were replaced by a unique sample containing the average value for the DNAm age. No duplicate healthy tissue DNAm measurements were found in the dataset.

Identification of patient-matched samples

In order to be able to make more rigorous comparisons between healthy and tumor tissues, a subset of the dataset was selected which only contained those tumor tissue samples for which a healthy counterpart was available. Since we had many more tumor samples than healthy ones, a lot of information from the original dataset was lost (only 87 out of the original 347 samples could be patient-matched). Therefore, we made use of the whole dataset in some parts of the analysis.

As for the expression data, only 37 tumor samples could be patient-matched and those were used as the input for the GSEA.

Creation of the input files for the GSEA

The software GSEA needs several files with a very specific format in order to perform the Gene Set Enrichment Analysis. In the present work we used the following input file formats:

1. Expression dataset (.gct)
2. Gene sets database (.gmt)
3. Phenotype labels (.cls)

The first was created by selecting the patient-matched tumor samples in the mRNAseq data file and doing some formatting: a first line had to be added containing the tag *#1.2*. Then a second line was added with the number of genes (20531) and samples (37) in the dataset. The third line only contains the headings of the data columns: *NAME*,

Description, and the identifier of each of the samples. The expression data begins in the fourth row. The first column contains the identifier of each gene, the second can either contain a description or be filled with NAs, and from the third column on the table contains the expression data.

The second is a gene set database which can either be downloaded or directly loaded when running the GSEA. We downloaded and used the gene set *c2.cp.v5.1.entrez.gmt*.This file required no preprocessing.

The third is a table that can either contain the classification of the samples in discrete categories or the values for one or more continuous variables. We created a file that contained the tag *#numeric* in the first line, and the values for the DNAm age of the tumor tissues and the acceleration according to expressions (1) and (2), each of them preceded by a line containing their respective tag: *#DNAmAge*, *#acc1*, *#acc2*.

Analysis of survival

In order to obtain the Kaplan-Meier plots shown in the *Results and Discussion* section, the R packages *survival* and *survminer* were used. In the analyses that did not require the DNAm age of the patient-matched healthy tissues, all of the samples in the dataset were used. The 87 sample patient-matched dataset was only used in the analyses were both the tumor and the healthy tissue DNAm age or the DNAm acceleration computed with the expression (2) were needed. For the sake of clarity, the dataset used in each of the plots is specified in the captions below the figures.

Gene Set Enrichment Analysis

We loaded the three files introduced in the previous section and performed a GSEA with the following parameter values:

- Collapse dataset to gene symbols: FALSE
- Number of permutations: 1000
- Permutation type: phenotype
- Enrichment statistic: weighted
- Metric for ranking genes: Pearson

The rest of the parameters were set to their default values.

GSEA takes as an input a matrix of gene sets that contains several gene sets. Given the gene set S and the ranking of genes L, the algorithm tries to discern whether the members of S are randomly distributed throughout L or primarily located at the top and at the bottom of the ranked list (Subramanian et al, 2005).

For that purpose, GSEA computes an Enrichment Score (ES), which is a measure of the overrepresentation of the set S at the top and the bottom of the ranked list L (Subramanian et al, 2005). In our case, the ES would reflect the overrepresentation of a given set of genes, for instance, the ones involved in the metabolism of polyamines, in those samples with highest or lowest DNAm age acceleration.

Then, the algorithm computes the statistical significance of the enrichment scores by comparing them to a null model built through random permutations on the phenotype. The ES is normalized for each gene so that the size of the gene set to which it belongs is taken into account. This yields a normalized enrichment score (NES). The False Discovery Rate (FDR) is then estimated by comparing the observed and the null distributions for the NES (Subramanian et al, 2005).

2. Results and Discussion

DNAm age prediction in healthy tissues

In his article, Horvath claims his predictor to be extremely accurate accross different healthy tissues and chronological ages, except for breast tissues, which he found to have their DNAm age accelerated. The Pearson correlation between the chronological age and the DNAm age in healthy tissues independent from the ones used in the training set was 0.96 according to his findings, and the median error was 3.6 years.

However, our study failed to replicate such results, since the correlation found between the DNAm age of healthy tissues and the chronological age of the patient from which each tissue was extracted was high (Spearman's correlation: 0.818, p-value=2.2e-16), yet quite lower than in the original results. The median error in our dataset was also much higher than the error reported by Horvath: 16.51 years. This means that 50% of our healthy samples had an absolute difference between the predicted DNAm age and the chronological age of the patient greater than 16.51 years.

Correlation between the DNAm age and the chronological age in tumor and healthy tissue samples

We plotted the DNA methylation age against the chronological age considering all the samples in the dataset.

We built two linear models that fitted the relationship between the DNAm age and the chronological age a) considering only the healthy samples and b) considering only the tumor samples. The blue and red lines correspond to the linear models that fit the healthy and tumor samples respectively (**Figure 1**).

We can observe that the points corresponding to the healthy samples appear to be very well aligned to the linear model. As we mentioned above, the correlation between chronological age and DNAm age in healthy tissues is very strong, yet not as strong as in Horvath's findings. Conversely, the correlation between the the DNAm age and the chronological age in tumor samples is much lower (rho = 0.335, p-value = 1.5e-10). We can also observe that the points are much more scattered than in the case of healthy tissues. We could thus conclude that, as it has been shown for many other cancers (Horvath, 2013), the DNA methylation clock is disrupted also in colorrectal adenocarcinomas.

In the same plot we included a straight line corresponding to the identity function (y = x). We can notice that both linear models (healthy and tumor) have a milder slope than what we should expect from a perfect age predictor. There are two possible explanations to this: either the DNAm clock underestimates the age of the healthy tissues or the age of the healthy colorrectal tissues is in fact decelerated with respect

to other tissue types. The latter explanation is supported by Horvath's finding (2013) that there are inter-organ variations in the DNAm age.

The Spearman correlation between the DNAm age of the tumor and the chronological age of the patient-matched samples (rho=0.404, p-value=1.06e-4), was slightly higher than the correlation between the DNAm age of the tumor and the DNAm age of the healthy tissues in those same samples (rho=0.355, p-value=1.59e-3).

An equivalent plot to the one in **Figure 1** was built using only the patient-matched samples and it looked almost identical to the one in **Figure 1**, but with fewer data points (not shown).

The **Table 1** summarizes all the information retrieved from this initial exploration. Note that the number of samples whose DNAm age is accelerated and decelerated correspond to the number of data points of each category falling above and below the identity line shown in **Figure 1**. We can observe that, when we look at the patient-matched samples, we cannot really say that there is a difference between healthy and tumor tissues regarding their DNAm age acceleration. Both healthy and tumor colorectal tissues appear to have their DNAm age decelerated with respect to the chronological age. It might be interesting to study whether there is a difference in DNAm acceleration of tumor tissues with respect to the healthy tissues, rather than with respect to the chronological age.

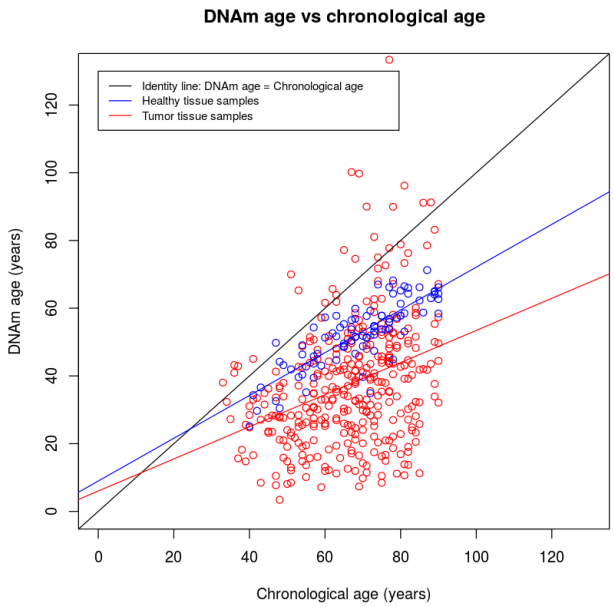

Figure 1. DNA methylation age vs chronological age of all the patients in the dataset. Regression lines for the healthy sample data points (blue) and for the tumor sample datapoints. The identity line (y=x) is shown in black.

	Healthy	Tumor (all)	Tumor (matched)
Accelerated DNAm age	1 (1.15%)	23 (6.63%)	4 (4.60%)
Decelerated DNAm age	86 (98.85%)	324 (93.37%)	83 (95.40%)
Cor(DNAm age, chrono. age)	0.818	0.335	0.404
p-value	2.2e-16	1.5e-10	1.10e-4

Table 1. For the three categories (healthy samples, all the tumor samples in the dataset, and patient-matched tumor samples): number and percentage of samples whose DNAm age was accelerated/decelerated, Spearman correlation between the DNAm age and the chronological age and p-value of such correlation.

DNAm age and the clinical variables

We computed the DNAm age acceleration using the two expressions mentioned in the introduction. After calculating the acceleration for each of the individual samples using the two formulas, we also calculated a "general" acceleration as the slope of the best-fit straight line in the plots DNAm age of tumoral tissues vs DNAm age of healthy tissues and the plot DNAm age of tumoral tissues vs chronological age (**Figure 2**). The aim of these linear regressions was to evaluate the overall behavior of the two different expressions of age acceleration, and this approach certainly misses the variations in acceleration across samples. Those variations and their correlation with clinical variables will be discussed later. In order to make the regressions comparable we only used the matched samples in both plots.

In **Figure 2** we can observe the lines corresponding to the two different expressions of DNAm age acceleration. The straight that best fitted the first expression of the acceleration is $y = 0.56x0.85$ and the second $y = 0.60x + 7.58$. We can therefore conclude that both expressions behave in a very similar way, as the slope of both straights is nearly identical and the two straights are almost parallel. In other words, the difference between the DNAm acceleration calculated using the two expressions is constant along the chronological age axis. This, however, does not mean that both expressions are equivalent, since one of them might be more sensitive to variations between individual samples.

Radiation

The dataset included information about whether each of the patients had undergone radiation therapy. We looked for differences in the distribution of the DNAm age and the DNA acceleration of patients who had been subject to radiation therapies vs those who had not. It is very hard to make comparisons between these two groups since, as we can observe in **Table 2**, only four subjects had undergone radiation therapy, while 68 had not. Since the sizes of the two groups are so

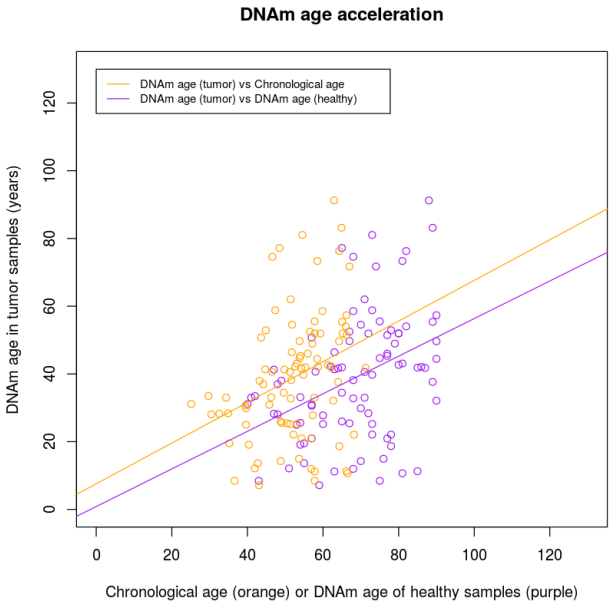

Figure 2. DNAm age of tumor tissues against a) chronological age (orange) and b) DNAm age of patient-matched healthy samples (purple).

different and one of them is so small, we cannot draw any conclusions on the possible correlations between radiation therapy and DNAm age acceleration.

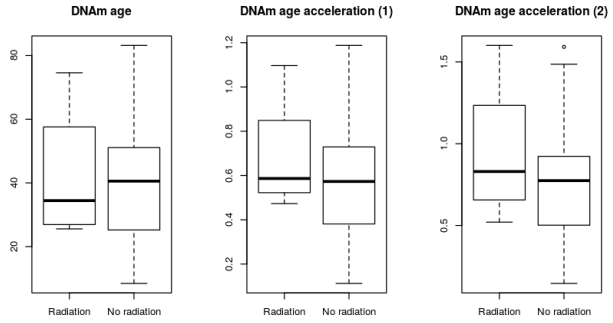

Figure 3. Boxplots showing the distribution of the DNAm age (left) and the DNAm age acceleration, calculated as in the expression (1) (middle) and according to expression (2) (right) in patients that had undergone radiation therapy and patients who had not. Only patient-matched samples were considered. Outliers are shown. The whiskers represent the most extreme data point which is no more than 1.5 times the interquartile range.

Tumor size

The T-stage measures the size and the extent of the primary tumor in a cancer. There are four main T-stages: T1, T2, T3 and T4. Each of the stages might be decomposed in substages in order to provide more details about the development of the tumor, but we only consider the four main stages here.

We studied the distribution of the DNAm age and the DNAm age acceleration. As we can see in **Table 2**, the group sizes are not even (there is only one sample in t1, but there are 63 in t3), which makes it difficult to make statistical comparisons between, for instance, the mean DNAm age acceleration per T-stage. However, it appears that the mean DNAm age acceleration increases with tumor size, although it is not statistically significant. A more complete dataset with even groups might shed some light on the link between tumor size and DNAm age acceleration. Horvath (2013) reported that only 4 out of 33 hypothesis tests led to nominally significant results when studying the correlation between tumor morphology/stage and DNAm age acceleration.

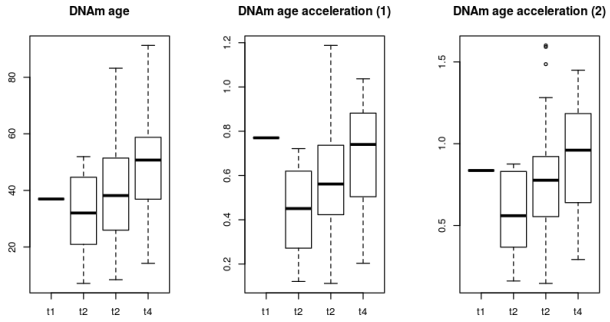

Figure 4. Boxplots showing the distribution of the DNAm age (left) and the DNAm age acceleration, calculated as in the expression (1) (middle) as in the expression (2) (right) according to their T stage. Only patient-matched samples were considered. Outliers are shown. The whiskers represent the most extreme data point which is no more than 1.5 times the interquartile range.

Hystological type

Four different hystological types were present in the dataset: colon adenocarcinoma, colon mucinous adenocarcinoma, rectal adenocarcinoma and rectal mucinous adenocarcinoma. As we can see in **Table 2**, the four categories are very unequally sized: we had access to 68 colon adenocarcinoma samples while only 2 rectal mucinous adenocarcinoma samples were available. Moreover, we see that the mean chronological age of the patients in each category is quite different, which makes it difficult to assess whether a given acceleration is due to the hystology-related grouping or to confounding variables like the chronological age. Since it is not pertinent to perform statistical tests on such unevenly sized and heterogeneous groups, we decided to at least explore the distribution of the variables of interest in each of the categories. In **Figure 5** we can see that the variance of the most populated category (CA) is much bigger than in the rest of the hystological types. We can also see that the distributions per hystological type of both expressions of the DNAm age acceleration are very similar, which suggests that both measurements are consistent with each other.

	Radiation		Tumor size (T-stage)				Hystological type				Tumor site		Sex	
	no	**yes**	**t1**	**t2**	**t3**	**t4**	**CA**	**CMA**	**RA**	**RMA**	**Col**	**Rec**	**M**	**F**
Number of samples	68	4	1	10	69	7	68	7	9	2	75	11	41	46
Mean Chronological age (years)	69.5	60.0	48.0	71.8	68.3	73.6	70.1	72.6	62.0	55.5	70.3	60.8	69.0	69.0
Mean DNAm age (years)	38.8	42.2	37.0	31.2	39.2	49.7	38.2	54.2	37.0	28.2	39.7	35.4	38.6	39.5
Mean DNAm age acc. (1)	0.56	0.69	0.77	0.43	0.57	0.68	0.54	0.76	0.62	0.50	0.56	0.60	0.56	0.58
Mean DNAm age acc. (2)	0.75	0.94	0.84	0.55	0.76	0.91	0.73	0.99	0.76	0.60	0.76	0.73	0.74	0.76

Table 2. Number of samples in each of the values that the categorical variables take: Radiation (no, yes); tumor size (t1, t2, t3, t4); hystological type of the tumor (CA: colon adenocarcinoma, CMA: colon mucinous adenocarcinoma, RA: rectal adenocarcinoma, RMA: rectal mucinous adenocarcinoma); tumor tissue site (Col: colon, Rec:rectum); sex of the patient (M: male, F: female). For each class, the number of samples falling in that category, the mean chronological age , the mean DNAm age, and the mean DNAm age acceleration according to expressions (1) and (2) are given.

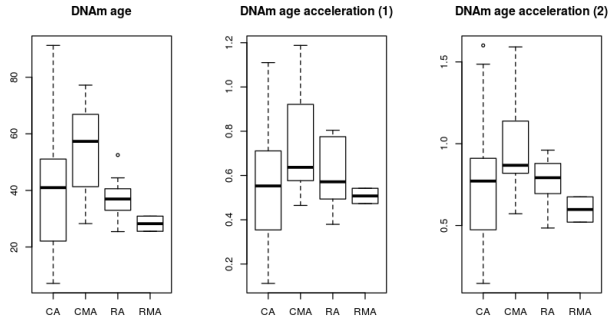

Figure 5. Boxplots showing the distribution of the DNAm age (left) and the DNAm age acceleration, calculated as in the expression (1) (middle) as in the expression (2) (right) according to their pathologic state. Only patient-matched samples were considered. Outliers are shown. The whiskers represent the most extreme data point which is no more than 1.5 times the interquartile range.

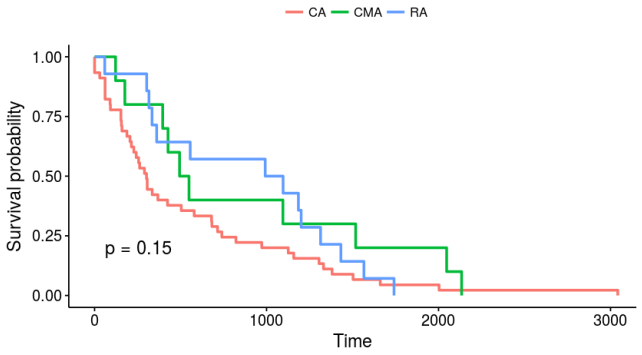

Figure 6. Kaplan-Meier curve of the patients according to the hystological type of their tumor (CA: colon adenocarcinoma, CMA: colon mucinous adenocarcinoma, RA: rectal adenocarcinoma). As there were too few patients with rectal mucinous adenocarcinoma, we could not build Kaplan-Meier curve for that category. Confidence intervals are not presented for the sake of clarity. The 347 patients were included in the survival analysis, although 278 observations could not be considered due to missing values.

Tumor tissue site

The clinical annotations included information about the location of the tumor. As the dataset contained samples of colorrectal cancers, some tumors were located in the rectum and some others in the colon. **Figure 8** shows the distributions of the DNAm age and the accelerations (1) and (2). These are not very informative since, again, the two categories have very different sample sizes and a difference of about ten years in their mean chronological age. However, it seems like there is no correlation between the location of the tumor and its DNAm age and DNAm age acceleration.

Sex of the patient

The boxplots in **Figure 7** show that there is no difference between males and females regarding the DNAm age of their tumors or their DNAm age accelerations (1) and (2). **Figure 8** shows the survival analysis of the patiens according to their sex. We can observe that there is no significant difference in survival between male and female patients.

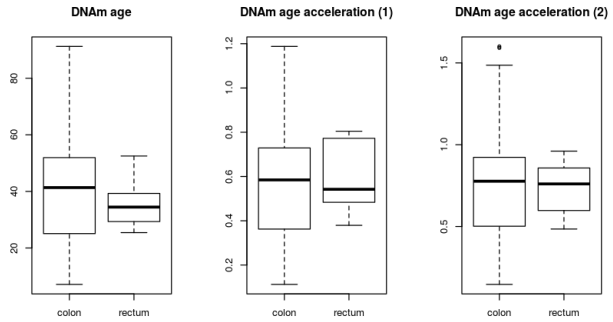

Figure 7. Boxplots showing the distribution of the DNAm age (left) and the DNAm age acceleration, calculated as in the expression (1) (middle) as in the expression (2) (right) according to the location of the tumor (colon, rectum). Only patient-matched samples were considered. Outliers are shown. The whiskers represent the most extreme data point which is no more than 1.5 times the interquartile range.

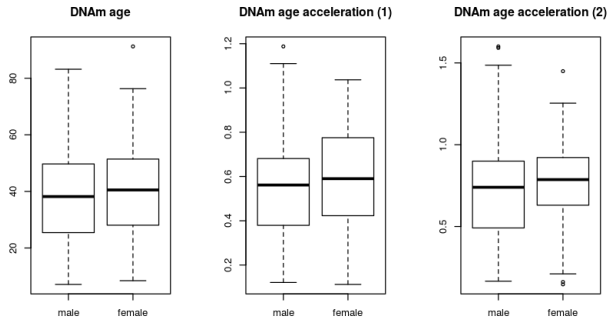

Figure 9. Boxplots showing the distribution of the DNAm age (left) and the DNAm age acceleration, calculated as in the expression (1) (middle) as in the expression (2) (right) according to the sex of the patient. Only patient-matched samples were considered. Outliers are shown. The whiskers represent the most extreme data point which is no more than 1.5 times the interquartile range.

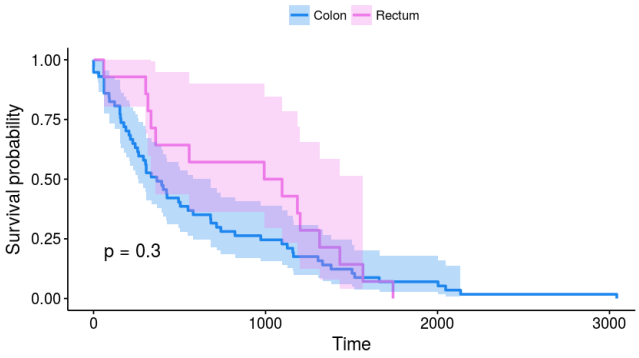

Figure 8. Kaplan-Meier curve of the patients according to the location of their tumor (colon or rectum). The shades represent the confidence intervals. 276 observations out of the original 347 could not be considered due to missing values.

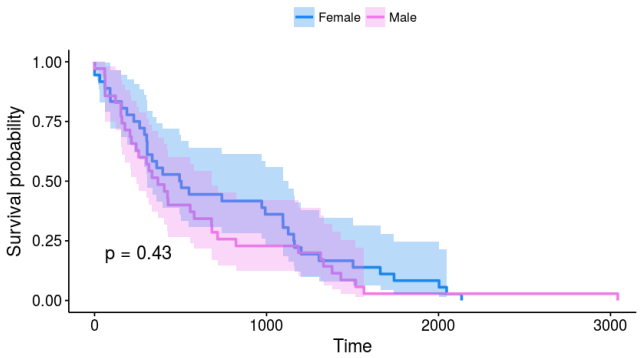

Figure 10. Kaplan-Meier curve of the patients according to their sex. The shades represent the confidence intervals. 276 observations out of the original 347 could not be considered due to missing values.

DNAm age of the tumor and patient survival

In order to explore the link between the DNAm age and the DNAm age acceleration of tumor tissues and the survival of the patients from which the tissues were extracted, we created evenly sized groups of patients according to the three variables (DNAm age, DNAm age acceleration (1) and DNAm age acceleration (2)) and studied their survival probabilities.

DNAm age of the tumor

We splitted the dataset in two groups: those with a tumor DNAm age below the median and those above. In **Figure 11** we can observe the Kaplan-Meier curves of the patients with younger and older-DNAm aged tumors. Of course, the plot must be interpreted with much caution since the classification does not take into account the chronological age of the patient. So, for instance, a 30-year-old patient with a 28-DNAm aged tumor might be classified in the younger category, while an 80-year-old patient with a 60-DNAm aged tumor might be classified in the older category, while the former has its DNAm

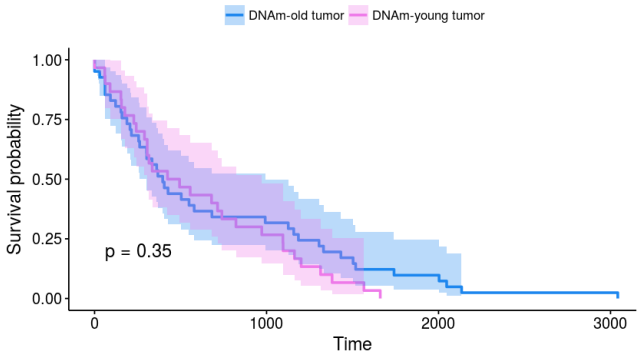

Figure 11. Kaplan-Meier curve of the patients split in two halfs according to their tumor DNAm age (young and old tumor). The shades represent the confidence intervals. 276 observations out of the original 347 could not be considered due to missing values.

age decelerated only by 2 years and the latter by 20 years. In other words, this classification aims to see the relationship between the absolute DNAm age of the tumors, rather than the relative up- or down-regulatation of the DNAm clock, and the survival. We can observe in the plot that both categories have a very similar survival curve, although it seems like those patients having DNAm-older tumors have sligtly better survival probabilities. Nevertheless, this difference is not statistically significant (p = 0.35).

DNAm age acceleration (1) of the tumor

Again, we splitted the dataset in two groups: the half with lowest and the half with the highest DNAm age acceleration according to expression (1). In **Figure 12** we can see the Kaplan-Meier plot of both groups. It appears that the patients whose tumors were only weakly decelerated (not so DNAm-young compared to their chronological age) had a better survival, yet this difference is not statistically significant (p-value = 0.47). Nevertheless, if this were the case, it would be consistent with the previous plot, since a higher DNAm age of the tumor (a lower deceleration of the DNAm age) would be related to a higher survival probability.

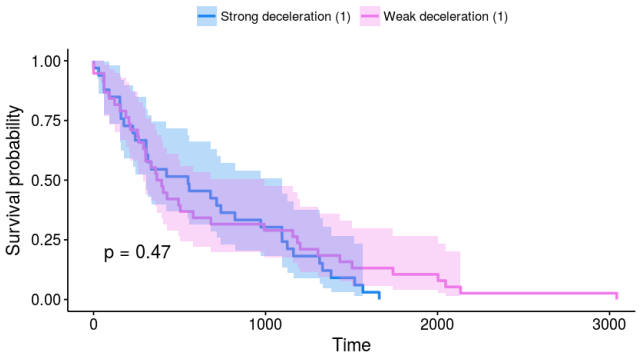

Figure 12. Kaplan-Meier curve of the patients according to DNAm age acceleration (1) of their tumor (strong deceleration: acceleration below the median DNAm acceleration; weak deceleration: above the median DNAm acceleration). The shades represent the confidence intervals. 276 observations out of the original 347 could not be considered due to missing values.

DNAm age acceleration (2) of the tumor

We did the same as in the previous section but using the expression (2) for DNAm age acceleration. As we need the DNAm age of the healthy tissues in order to compute the acceleration in this way, we could not use the whole dataset but only the samples from which we had all the necessaary information. This is why we can appreciate in **Figure 13** that the plot is coarser than the one in **Figure 12**.

Gene Set Enrichment Analysis

DNAm age as a covariate

The upregulation of 581 gene sets out of 1071 was found to be related to an increase in DNAm age, although none of them was significantly upregulated at a FDR $< 25\%$. One gene set was significantly enriched at a nominal p-value < 0.01 (REACTOME METABOLISM OF POLYAMINES: genes involved in the metabolism of polyamines). **Figure 14** shows the enrichment plot of this gene set.

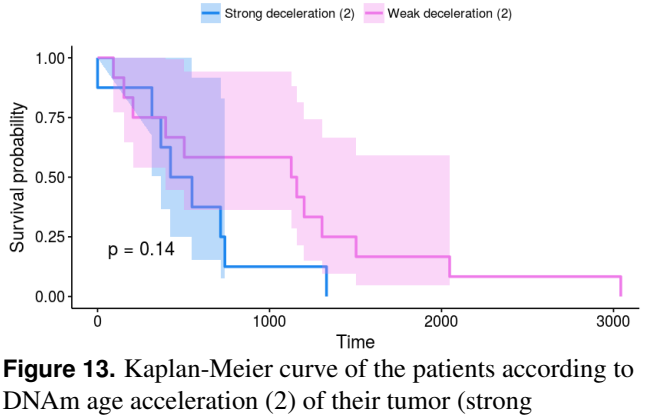

Figure 13. Kaplan-Meier curve of the patients according to DNAm age acceleration (2) of their tumor (strong deceleration: acceleration below the median DNAm acceleration; weak deceleration: above the median DNAm acceleration). The shades represent the confidence intervals. Only the patient-matched samples were used. 67 observations out of the original 87 could not be considered due to missing values.

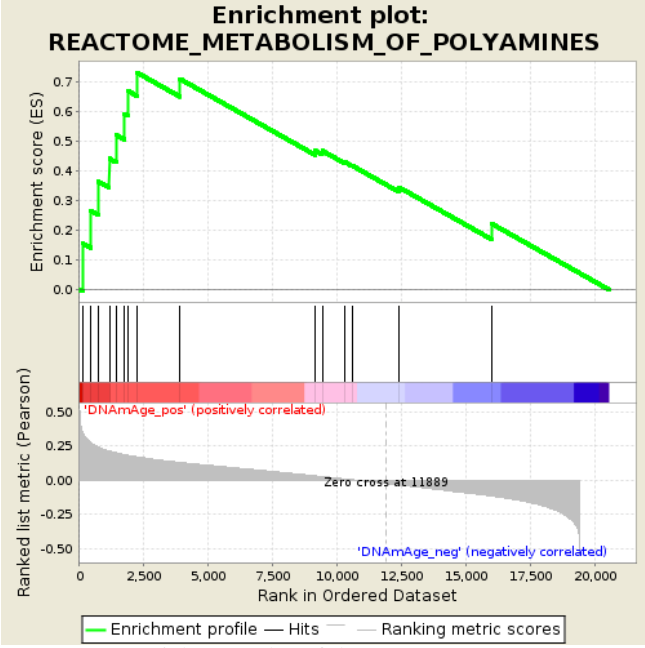

Figure 14. Enrichment plot of the gene set Reactome Metabolism of Polyamines.

490 gene sets were found to be negatively correlated with DNAm age, and one of them was significantly enriched at $FDR < 25\%$ (REACTOME AMINE DERIVED HORMONES: set of genes involved in amine-derived hormones). Two gene sets were found to be significantly negatively corre-

lated with DNAm age: REACTOME PTM GAMMA CARBOXYLATION HYPUSINE FORMATION AND ARYLSULFATASE ACTIVATION (set of genes involved in PTM: gamma carboxylation, hypusine formation and arylsulfatase activation) and REACTOME FORMATION OF FIBRIN CLOT CLOTTING CASCADE (set of genes involved in the formation of fibrin clot in the clotting cascade).

In **Figure 15** we can observe the evolution of the enrichment score for the amine derived hormone set, and how it became more negative as the algorithm walked down the list L, meaning that the members of that gene set were most likely to appear at the bottom of the ranked list, that is to say, they are negatively correlated with the DNAm age.

Figure 16 shows a very similar plot in which we can also appreaciate the negative correlation between the DNAm age and the expression of the set of genes involved in PTM: gamma carboxylation, hypusine formation and arylsulfatase activation. If we compare **Figures 15** and **16**, we can observe that the former has most of the hits in the blue region of the plot, meaning that the negative correlation betwen the expression of that set of genes and the DNAm age is much stronger than the one in **Figure 16**. The ES score decreases very fast as the algorithm goes down through the ranking and there are no hits in the red region, that is to say, no gene belonging to that gene set is found to be upregulated in the samples whose DNAm age is very high. Conversely, in **Figure 16** we see that some of the genes belonging to the set under study are found to be upregulated in samples that have a medium-high DNAm age, but the much higher number of genes in the set that appear to be overrepresented towards the end of the ranking (the leading edge subset of genes) compensate for this, yielding a statistically significant negative correlation. The enrichment scores for the rest of the results are not included in the report as they look very similar to these and are not particularly informative.

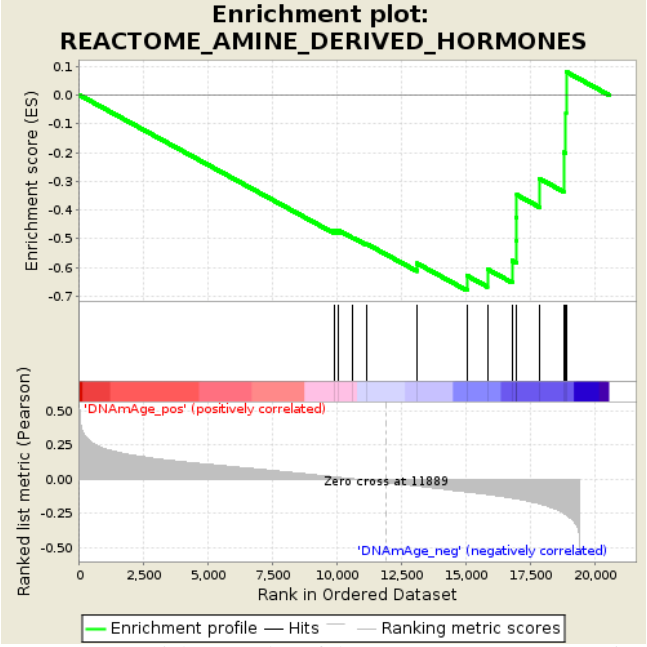

Figure 15. Enrichment plot of the gene set Reactome Amine Derived Hormones.

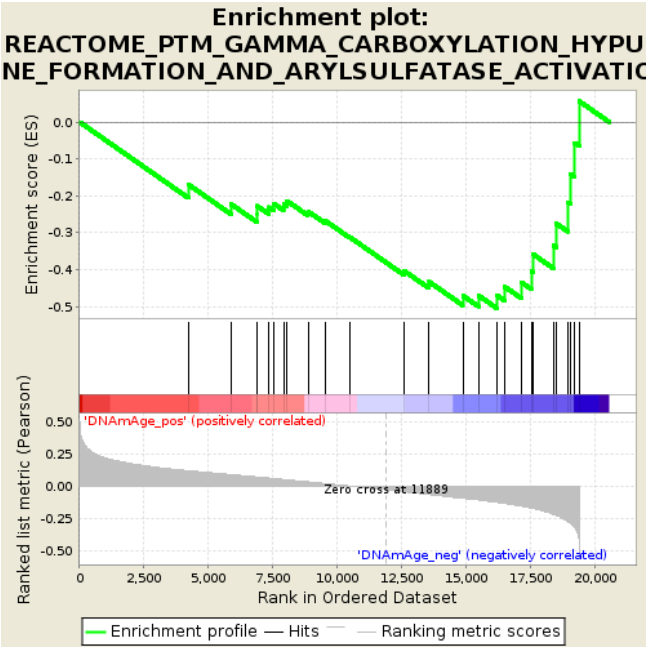

Figure 16. Enrichment plot of the gene set Reactome PTM Gamma Carboxylation Hypusine Formation and Arylsulfatase Activation.

DNAm age acceleration (1) as a covariate

602 gene sets were positively correlated and 469 gene sets were negatively correlated with the DNAm acceleration (1), although none of those results were statistically significant at the $FDR < 25\%$ level. Only two gene sets were found to have a correlation with a nominal p-value < 1: KEGG HYPERTROPHIC CARDIOMYOPATHY HCM (genes involved in HMC, a inherited cardiac disorder), which was positively correlated with DNAm age acceleration, and REACTOME GLYCOSPHINGOLIPID METABOLISM (genes involved in glycosphingolipid metabolism), which was negatively correlated with DNAm age acceleration.

DNAm age acceleration (2) as a covariate

500 gene sets were positively correlated and 571 negatively correlated with the DNAm acceleration (2), but, as with the expression (1) of the acceleration, none of them were significantly correlated with $FDR < 25\%$. Only one gene set was found to be significantly negatively correlated with the DNAm age acceleration (2): ST TUMOR NECROSIS FACTOR PATHWAY (genes belonging to the tumor necrosis factor pathway). It is interesting to note that the gene sets that showed significant correlation with the DNAm age acceleration were different depending on the expression of the acceleration that we used. This suggests that it is not that clear and that the correlations found are biologically relevant or at least truly related to the covariates that we were trying to

study, and that we should be cautious when drawing conclusions from this GSEA results. Moreover, the fact that many of the gene sets for which a significant correlation was found in terms of nominal p-value did not meet the FDR $<25\%$ reguirement, which means that those results might have been considered positive and reliable in a different setting, even though they should be regarded with much caution.

3. Conclusions

Even though the present work did not yield many positive results, we could draw some some conclusions regarding the methodology we used.

The first one is that, although it might seem more to use the DNAm age of the healthy tumors in the calculation of the acceleration, the loss of valuable information resulting from the selection of patient-matched samples is a bigger issue than it might seem at first. In our case, the dataset was drastically reduced (only 87 out of the 347 tumor samples had a healthy sample counterpart), which decreased the statistical power of the tests performed. In general, I would suggest calculating the DNAm acceleration as the ratio against the chronological age, since the chronological age is usually available for all the samples and is very easy to retrieve.

As for the GSEA, we can conclude that it is a very useful tool to analyse expression data and that it facilitates the assimilation of large amounts of results that would otherwise be hard to integrate. It is important that we analyse critically the positive results obtained in a GSEA, since, as its developers explain in their article, the the main goal of the GSEA is to generate hypotheses (Subramanian et al, 2005), which is the reason why they decided to replace the too conservative measure of the familywise-error rate (FWER) by the more loose criterion of the FDR. Even though it is important that a tool like GSEA enables us to generate new hypotheses, it becomes more important in such a setting that researchers interpret the results with a very critical mindset.

4. Bibliography

Fernandes Durso D, Bacalini MG, Faria do Valle I, Pirazzini C, Bonafe M, Castellani G, Caetano Faria AM, Franceschi C, Garagnani P, Nardini C, *Aberrant Methylation patterns in colorectal cancer: a meta-analysis*, Oncotarget, 2017, Vol. 8 (NO. 8), pp: 12820-12830.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gilette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP, *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*, PNAS, 2005, vol.102, (NO. 43), 15545-15550.

Horvath S, *DNA methylation age of human tissues and cell types*, Genome Biology, 2013, vol. 14, (NO. 10), 14:R115.