DNA methylation age in LUSC cancer

BINF-F401: Functional genomics

Nil Fernandez Lojo

June 2, 2020

Abstract

This reports presents a study of the DNA methylation age in lung squamous cell carcinoma (LUSC) tumour cells. The analysis was performed on the publicly available data from the TCGA project. A statistically significant decrease in tumour DNA methylation age was observed in tumour cells however no statistically significant correlations between this decrease and the patient's clinical variables were found. Moreover, using RNAseq data from the tumours, it was found that some genes sets from the MSigDb C2:CP dataset that code for proteins linked to lung cancer had their expression correlated with the decrease in DNA methylation age. Finally, a small, but statistically significant, negative correlation between the DNA methylation age and both the number of somatic mutations and the number of genes with abnormal copy number in tumour cells was observed.

Introduction

Several studies showed that DNA methylation of several genes in animals varies with their age (e.g. [1]). Out of the 4 bases of DNA, adenine and cytosine can be methylated but in mammals, DNA methylation mainly occurs on a cytosine followed by a guanine in the 5'→3' direction (these sites are called CpGs).

Steve Horvath showed in several papers that the age of an animal can be predicted with a high accuracy using a weighted sum of the methylation of some specific CpG [2, 3]. In [2] (and its erratum [4]), he develops an age predictor (called DNAm age) for humans based on 353 CpGs where 193 are positively correlated with age and 160 negatively correlated. To develop it, he used published data consisting of 7884 non cancer samples of 82 individuals from 51 different tissue and cell types. His age predictor on the test dataset (samples from 32 out of the 82 individuals) has a median error of 3.6 years.

Steve Horvath also studied the correlation of DNAm age and chronological age in human tumour cells in [2] which contained numerical errors that were corrected in [4]. It was shown that DNAm age was reduced in some cancers while increased in others.

This report contains an analysis of the relationship of DNAm age with clinical variables and other genomic and transcriptomic data from patients with lung squamous cell carcinoma (LUSC). Lung squamous cell carcinoma is the second most common type of lung cancer and it is heavily correlated with tobacco smoking history of patients [5].

The rest of this paper is structured as follows: in section 1, the correlation of DNAm age with chronological age in healthy and LUSC tumour cells is presented. Then in section 2, the correlation of DNAm age of tumour samples with patient's clinical variables is studied. In section 3, the correlation of gene expressions in LUSC samples with methylation age is analysed. Then in section 4, the correlation of DNAm age with the number of somatic mutations and the copy number variations (cnv) of genes is studied.

All the data used for this project comes from the cancer genome atlas (TCGA) other than the methylation age of the samples that was precomputed by Vincent Detours. The data was analysed on R and the scripts, package versions and the source data are available on `https://github.com/Nil-Fernandez-Lojo/DNAm_age_LUSC`.

	Estimate	Standard error
slope	0.64	0.06
Intercept	13.3	4.22

Table 1: Standard deviation of the inferred parameters from equation 1.

	Estimate	Standard error
slope chronological age	0.54	0.11
slope sample type	-17.6	2.0
Intercept	19.9	7.5

Table 2: Standard deviation of the inferred parameters from equation 2.

1 Healthy versus tumour cells

The methylation age of 350 cancer samples and 69 healthy samples was provided. In figure 1, the DNAm age of healthy samples is compared to the chronological age of the patients. As expected they are very linearly correlated : Spearman $r = 0.696$ with an estimated p value[a] of 3×10^{-11} and a Pearson $r = 0.786$ with a p value of 10^{-15}. Surprisingly the DNAm age is consistently lower than the chronological age. An ordinary least squares linear regression was performed on the healthy samples data and the best fit was

$$\hat{a}_{\text{DNAm healthy}} = 13.3 + 0.64 a_{\text{chron}} \tag{1}$$

where $\hat{a}_{\text{DNAm healthy}}$ is the predicted DNAm age and a_{chron} is the chronological age. The standard deviations of the 2 inferred parameters are given in table 1.

An intercept of 0 and a slope of 1 are not in the 95% confidence interval of the regression coefficient estimation. In [2], DNAm age is an unbiased estimator of chronological age, therefore either a different normalisation was used in the computation of the given DNAm age or the DNAm age of healthy lung cells is lower than other cell types.

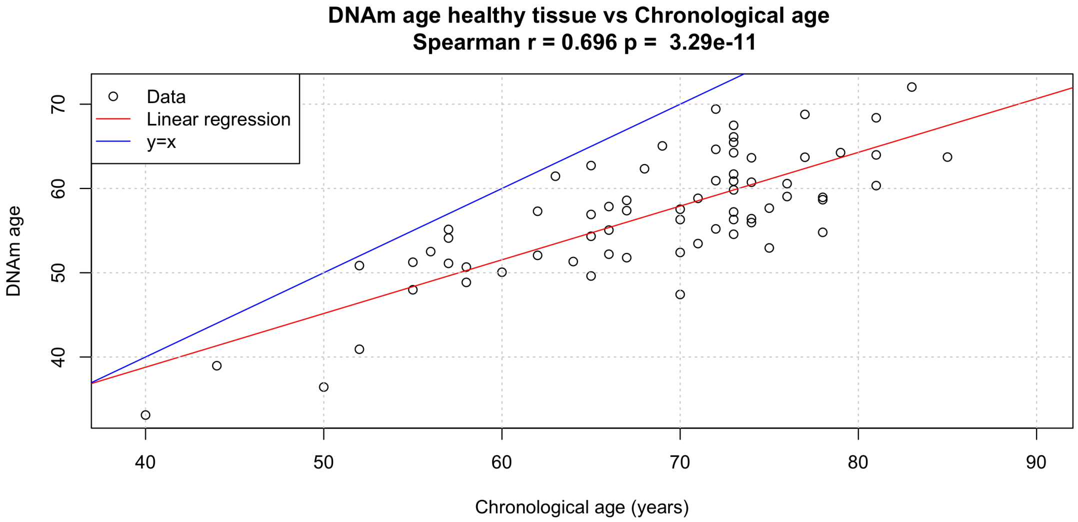

Figure 1: Scatter plot of DNAm age of healthy samples versus their chronological age. In blue, it is the identity line and in red, the linear model 1 prediction.

[a] The p-value of the Spearman estimator is defined as the probability that the Spearman coefficient has a more extreme value if the rank of the samples are shuffled at random. While this is can be computed exactly for less than 8 samples (it would require 8! permutations), it needs to be approximated when there are more samples. The given p-values (obtained using the cor.test() function from R) are obtained using the AS 89 approximation algorithm that takes a simple normal approximation [6]. However it assumes no repeated values which is not the case in the age of the patients. A better estimation of the p value could be obtained with a Monte Carlo estimation of the p value by sampling permutations at random and computing the fraction of permutations that have a more extreme Spearman coefficients. In the rest of this report, all the given p values for Spearman coefficients are computed with the AS 89 algorithm.

In figure 2, the DNAm age of tumour samples is compared to the chronological age of the patients. While the tumour DNAm age is still correlated with the chronological age (Spearman $r = 0.29$ with an estimated p value of 2×10^{-6}), the correlation is weaker and the DNAm age of the tumours is in average lower than the age predicted by the linear model 1. This indicates that the DNAm age of LUSC cancer tumour cell is decreased. This can also be observed in figure 3 where the DNAm age is compared between tumour and healthy lung tissues. The DNAm age of tumour cells is correlated to the DNAm age of healthy samples (Spearman $r = 0.30$ with an estimated p value of 0.012) but is consistently lower than it.

To confirm the statistical significance of the decrease of DNAm age in tumour cells, a linear model of the DNAm age on the chronological age and sample type was computed using ordinary least squares. The best fit was

$$\hat{a}_{\mathrm{DNAm}} = 19.9 + 0.54 a_{\mathrm{chron}} - 17.6t \tag{2}$$

where $t = 0$ for healthy tissues and $t = 1$ for tumour tissues. The standard deviation for these parameters are given in table 2. The DNAm age decrease due to cancer is significant (more than 8 standard deviations away from 0). To fit this model, data from patients who had both their DNAm age measured in tumour and healthy cells was used (in total 69 patients) to avoid an over representation of tumour samples.

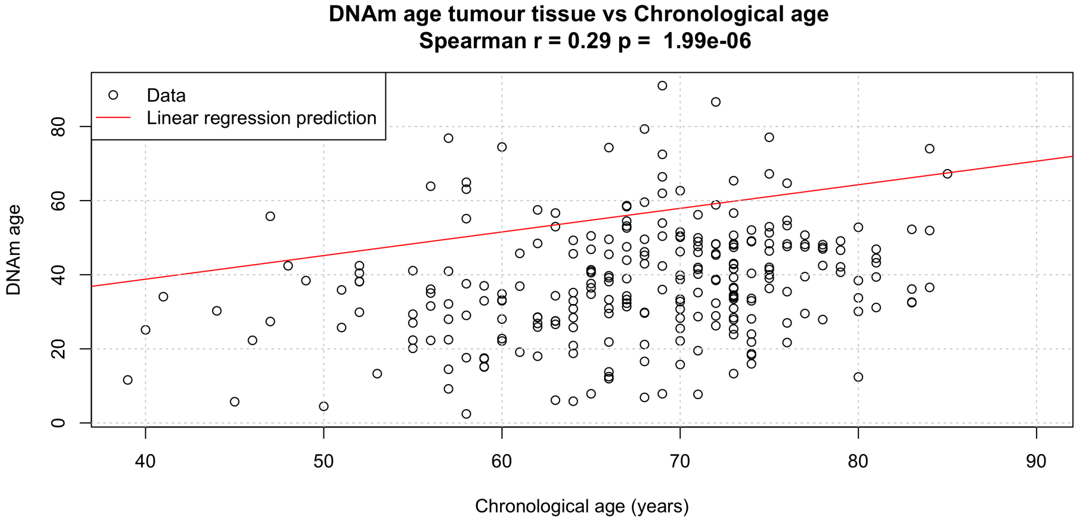

Figure 2: Scatter plot of DNAm age of tumour samples versus their chronological age. The red line is the linear model 1 prediction.

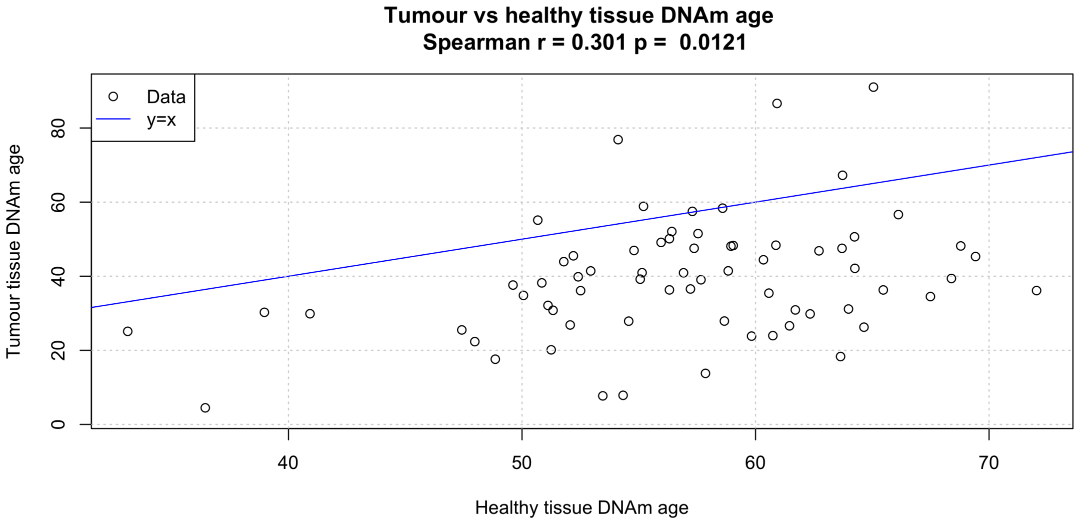

Figure 3: Scatter plot of DNAm age of tumour samples versus healthy samples. Each point in the scatter plot corresponds to the DNAm age of the same patient on a tumour and healthy cell. In blue is the identity line.

Multiple metrics could be used to measure the decrease of DNAm age in tumour cells independently of the chronological age of the patient. One possibility is to subtract (or divide) the tumour cell DNAm age by the healthy cell DNAm age of the patient. However, only 69 patients had their DNAm age measured in healthy cells out of the 350 patients who had their DNAm age measured in tumour cells. Therefore more than 75% of the samples would be excluded from the analysis. A better option is the subtract (or divide) the DNAm age of the tumour samples by the DNAm age predicted by the linear model from equation 1. This avoids the problem of sample reduction. Dividing (or subtracting) the tumour DNAm age by the chronological age of the patient is not a good option since the intercept term in the model from equation 1 is statistically significantly different from 0 (equivantly for the subtraction, the slope in the model from equation 1 is statistically significantly different from 1). In [2], the measure for difference in DNAm age and chronological age is called DNAm acceleration and is defined as the difference between DNAm age and chronological age. This is only meaningful since the two quantities have the same scale. To overcome this difference in scale, it was decided to define the DNAm acceleration (which is technically not an acceleration) as the difference between the DNAm age and the DNAm age predicted in equation 1 i.e.

$$\text{Acceleration} = a_{\text{DNAm}} - \hat{a}_{\text{DNAm healthy}} \tag{3}$$

In figure 4, it can be seen that the DNAm acceleration is uncorrelated with the chronological age.

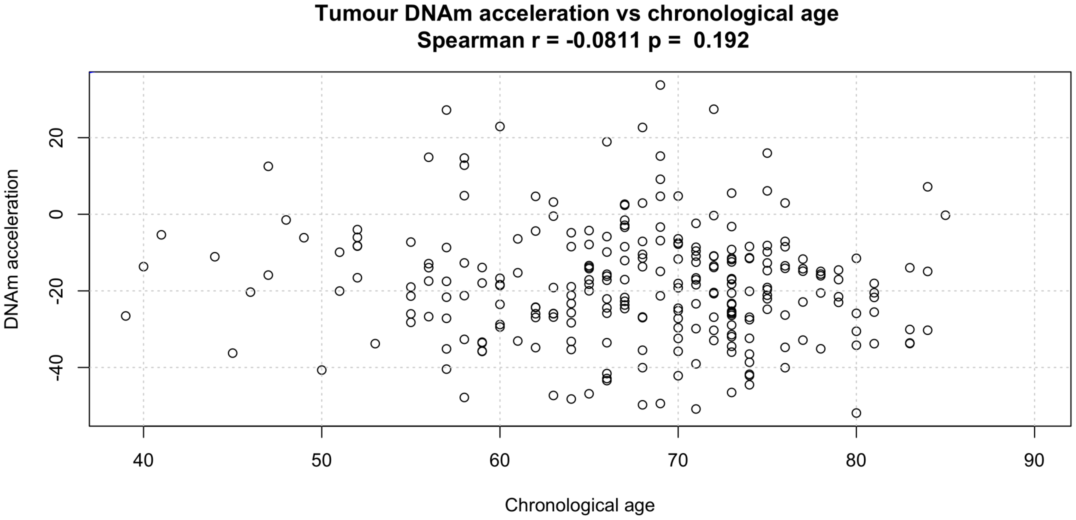

Figure 4: Scatter plot of DNAm acceleration of tumour samples versus their chronological age.

2 DNAm age versus clinical variables from patients with LUSC

In this section, the pairwise correlations of tumour DNAm acceleration and age with clinical variables are computed. While it is more pertinent to compare the DNAm acceleration with the clinical variables, it was also required for this assignment to compare the DNAm tumour age with them. The correlation for numerical clinical variables is measured with the Spearman coefficient. For categorical clinical variables, it is computed with ANOVA tests or variations of it if the ANOVA required assumptions are violated. Finally for survival data, the correlation is computed using a Cox analysis. The considered clinical variables are given in table 5. The pathologic stages could have been considered as numerical variables since they are ordered but it was decided to consider them as categorical in case there were non monotonously increasing or decreasing correlations with the DNAm acceleration or age. Moreover for some patients, the pathologic substages were given (e.g. T stage 1a or 1b) while for other patients they were not (e.g. T stage 1), therefore it was decided to merge all the substages (e.g. T stage 1a and 1b mapped to T stage 1). Since the chronological age was missing for some patients, fewer samples could be given a DNAm acceleration (350 samples for DNAm tumour age and 260 for DNAm tumour acceleration).

To control the family false positive error rate, a Sidak correction was applied to the significance p value threshold required to consider the correlation as statistically significant. The Sidak correction is given by equation

$$\alpha_{\text{Sidak}} = 1 - (1 - \alpha)^{\frac{1}{m}} \tag{4}$$

where α is the original significance threshold and m is the number of comparisons made. It is less conservative than the Bonferroni correction and is exact (i.e. the probability that there is at least 1 false positive is α) if the data for each pairwise tests are independent. However the Sidak correction is still too conservative if the data is not independent.

The results of the pairwise correlation tests are given in table 3. It was decided to use a statistical significance threshold of $\alpha = 0.05$ and 14[b] comparisons were made therefore the corrected

[b] Technically 28 comparisons are made, 14 for age and 14 for acceleration. However since it is more pertinent to analyse the acceleration than the age, the correlation's results for age are ignored and just added for reference (since it was required in the coursework description).

significance threshold is 0.0037. Therefore none of the correlations were significant. Let's note that 2 clinical variables (pathologic stage and number of packs smoked) are functions of other clinical variables and other clinical variables are likely dependent, therefore the Sidak correction may be slightly too conservative. The correlation between DNAm tumour acceleration and race is still probably not significant since a post hoc Tukey's range test showed that the intergroup difference was only significant for the Asian race. There are only four samples for the Asian race and one of them (DNAm acceleration of participant 3407) seems to be an outlier, it is the DNAm acceleration sample with maximum value out of the whole dataset. If that sample is removed, the p value for the ANOVA test becomes 0.15. The normality and variance homogeneity conditions for ANOVA like tests are given in appendix A. In conclusion no statistical significance between tumour DNAm acceleration and clinical variables was found. If a significant correlation was found, the same analysis should have been performed for healthy cells to ensure that the correlation is only present in tumour cells and not in all cells. In [4], a correlation test between DNAm tumour acceleration in LUSC samples and tumour stage was performed and not statistically significant correlation was found either.

To have a better overview of the different dependencies, an overall (rather than a pairwise) analysis of the correlations would be interesting. For instance the DNAm acceleration could be modelled as a linear function of the clinical variables (after one hot encoding the categorical variables), unfortunately, there are many problems with that approach: Firstly there are a multiple missing values for some variables which decreases the overall number of available samples. Then, there is a risk of overfitting due to a small number of samples and a large number of inferred parameters. Moreover, it assumes a linear dependence between categorical variables and DNAm acceleration. Finally some clinical variables are correlated (e.g. the different cancer pathologic stages are correlated), which decreases the statistical power of the linear model.

A survival analysis of the patients on the DNAm tumour age and acceleration was performed on R with the Survival package. The dependency of the patient's survival on the DNAm acceleration was first visually explored by plotting a Kaplan–Meier of the patients with cancer with high DNAm acceleration (more than the mean DNAm acceleration of tumour samples) or with low DNAm acceleration (less or equal to the mean DNAm acceleration of tumour samples) in figure 5. Days to last follow up were used to right censor the data. Days to last known alive were also given in the TCGA database but for fewer samples than days to last follow up (68 versus 283). In figure 5, no large difference between the survival of patients with high and low DNAm acceleration can be visually observed. To rigorously analyse the dependence of the patient's survival on the DNAm tumour acceleration, a Cox proportional-hazards model regression was performed. It models the hazard function as:

$$h(t) = h_0(t)e^{\beta \times \text{DNAm acceleration}} \tag{5}$$

However the inferred parameter β in the hazard function was not statistically significantly different from 0 as shown in table 4. This analysis was also performed for the tumour DNAm age but did not give statistically significant results either.

In conclusion, no statistically significant correlation was found between DNAm acceleration in tumour cells and clinical variables.

Variable	**DNAm tumour age**		**DNAm tumour age**	
	Test	p **Value**	**Test**	p **Value**
Karnofsky score	Spearman	0.89	Spearman	0.44
Number of years smoked	Spearman	0.03	Spearman	0.21
Number of packs per year smoked	Spearman	0.38	Spearman	0.25
Number of packs smoked	Spearman	0.55	Spearman	0.59
T stage	Welch's ANOVA	0.03	Kruskal-Wallis	0.37
N stage	Kruskal-Wallis	0.22	Kruskal-Wallis	0.36
M stage	Kruskal-Wallis	0.02	Kruskal-Wallis	0.02
Pathologic stage	Kruskal-Wallis	0.06	Kruskal-Wallis	0.10
Gender	Kruskal-Wallis	0.03	Kruskal-Wallis	0.01
Radiation therapy	ANOVA	0.76	Kruskal-Wallis	0.29
Histological type	Kruskal-Wallis	0.62	Kruskal-Wallis	0.91
Residual tumour	Kruskal-Wallis	0.79	Kruskal-Wallis	0.89
Ethnicity	ANOVA	0.61	ANOVA	0.74
Race	ANOVA	0.005	ANOVA	0.007

Table 3: Correlation significance of clinical variables with DNAm tumour age and acceleration. It was decided to use a 5% p-value threshold to control the false positive errors. After applying the Sidak correction, the threshold for significance is $\alpha = 1 - (1 - 0.05)^{1/14} = 0.0037$. Therefore none of the correlations are considered statistically significant.

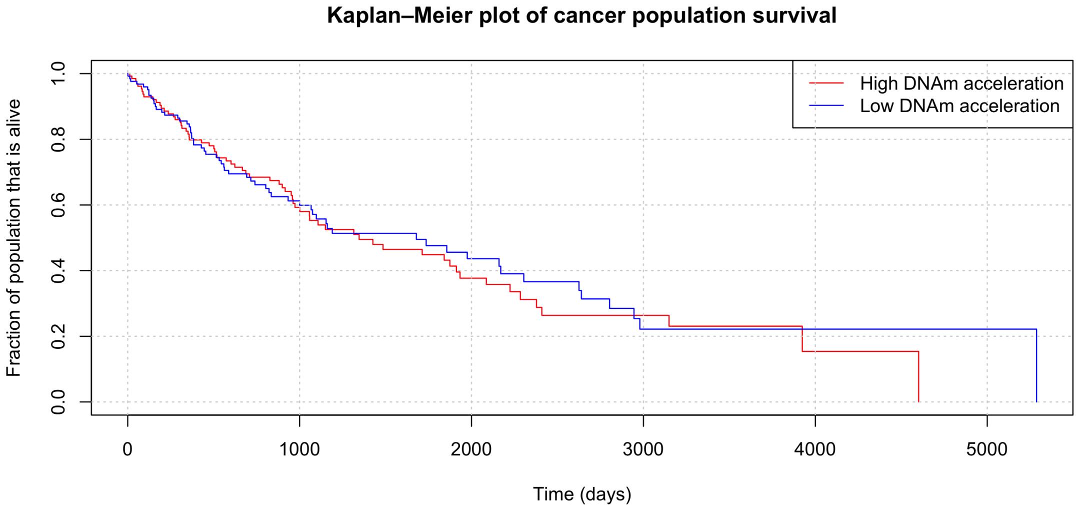

Figure 5: Kaplan–Meier plot of cancer population survival. The population having a high DNAm acceleration (more than the mean DNAm acceleration in tumour samples) is represented in red while the one with low DNAm acceleration (less or equal to the mean DNAm acceleration of tumour samples) is represented in blue.

	$\hat{\beta}$	p-value
DNAm acceleration	0.002	0.762
DNAm age	0.007	0.22

Table 4: Cox analysis results

Variable	**type (number of samples)**	**Description**
Karnofsky score	Numerical (93)	Measure of general wellbeing of the patient. It varies between 100 (no symptoms of cancer) and 0 (dead). It was initially designed to determine the patient's ability to survive chemotherapy.
Number of years smoked	Numerical (174)	Number of years between smoking onset and year of cancer diagnosis
Number of packs per year smoked	Numerical (227)	Average number of packs smoked per year
Number of packs smoked	Numerical (167)	Number of years smoked times number of packs per year smoked
T stage	Categorical: • stage 1 (60) • stage 2 (158) • stage 3 (27) • stage 4 (15)	Tumour size [7]: • stage 1 : size < 3cm • stage 2 : 3cm < size < 5cm • stage 3 : 5cm < size < 7cm • stage 4 : 7cm < size
M stage	Categorical: • stage 0 (221) • stage 1 (6)	No distant metastasis (stage 0), distant metastasis (stage 1)
N stage	Categorical: • stage 0 (163) • stage 1 (69) • stage 2 (21) • stage 3 (3)	Cancer cells in lymph nodes [7]: • stage 0: No lymph node metastasis • stage 1: lymph node metastasis in lung or hilum • stage 2: in mediastinum or subcarinal lymph node metastasis • stage 3: more distant lymph node metastasis
Pathologic stage	Categorical: • stage 1 (135) • stage 2 (70) • stage 3 (46) • stage 4 (6)	Function of TNM stages [7]: • stage 1 : N=0, M=0 and T <2.b • stage 2 : not stage 1 and (N=0, M=0 and T<4) or (N=1 M=0 and T<3) • stage 3 : not stage 2 and M=0 • stage 4 : M=1
Gender	Categorical: • female (64) • male (196)	
Radiation therapy	Categorical: • no (197) • yes (21)	
Histological type	• basaloid squamous cell (6) • papillary squamous cell (3)	Cell type in which the cancer initiated.
Residual tumour	Categorical: • r0 (207) • r1 (4) • r2 (2)	Size of residual tumour [8]: • r0 : No residual tumour • r1 : Microscopic residual tumour • r2 : Macroscopic residual tumour
Ethnicity	Categorical: • hispanic or latino (7) • not hispanic nor latino (158)	
Race	Categorical: • Asian (4) • Black or african american (10) • White (188)	
Vital status	Catgorical, survival: • alive (132) • dead (128)	
Days to death	Numerical, survival (127)	If vital status = dead, number of days elapsed between the time the clinical and multi-omics data was measured and the time of death. If vital status = alive, it is set to NA.
Days to last follow up	Numerical, survival (132)	If vital status = alive, number of days elapsed between the time the clinical and multi-omics data was measured and the last follow up with the patient. If vital status = dead, it is set to NA.

Table 5: Description of the clinical variables used. The number of samples given in the middle column is the number of samples used for the correlation with the DNAm acceleration of tumour. The number of samples used for the correlation for the tumour DNAm age is a bit higher since the chronological age of some patients is missing in the dataset.

3 DNAm age versus gene expressions in LUSC samples

In this section, the correlation between gene expression of tumour cells with their DNAm acceleration is investigated with the limma R package. The RNAseq data used in this report and obtained from the TCGA database is a read count of each gene that was generated with the RSEM algorithm [9]. RSEM first maps the raw reads to the human reference genome with bowtie or other read alignment software and then postprocesses the ambiguous maps with the expectation maximisation algorithm. For this report the **non normalised** RSEM counts were used since some limma functions require the usage of non normalised counts. The details of the implementation of the described analysis in this section can be found in `https://github.com/Nil-Fernandez-Lojo/DNAm_age_LUSC`.

The correlation between gene expression of tumour cells with their DNAm acceleration is analysed in two parts. Firstly, the correlation between tumour DNAm acceleration and expressions of the canonical pathways (CP) gene sets in the curated genes sets (C2) from the MSigDB database is computed with the limma camera function [10]. The C2:CP gene sets are curated from online pathway databases that include among others BioCarta, KEGG and Reactome. Secondly the correlation between individual gene expressions and DNAm tumour age is computed with the general limma procedure. These analyses are also performed on the healthy tissues to check if the found correlations in the tumour cells are also present in the healthy cells. However fewer healthy samples had their RNAseq profile and DNAm methylation measured : 8 for healthy samples and 259 for tumour samples.

Before computing the correlations, the RSEM count data is first filtered by removing reads with low counts with the filterByExpr function from the edgeR package [11]. Then a normalisation of the library sizes is then computed with the TMM method [12]. For all correlations computed with the limma functions, the read counts (the CPM actually) need to be normally distributed across different samples and have an homogeneous variance across different genes independently of their mean counts per million. The voom function transforms the data to ensure that these 2 properties are fulfilled.

With the limma package, the gene expression was modelled with the linear model:

$$A_{i,j} = b_i + w_{i,\text{acc}} \times \text{acc}_j + w_{i,\text{chron age}} \times \text{age}_j \tag{6}$$

where $A_{i,j}$ is the Voom normalised count of gene i in patient j; b_i , $w_{i,\text{acc}}$ and $w_{i,\text{chron age}}$ are parameters of the model, acc_j is the tumour DNAm acceleration of patient j defined in equation 3, and age_j is the chronological age of patient j.

The C2:CP genes sets expression that were found to be significantly correlated by the camera algorithm with the DNAm acceleration are given in table 6 for tumour samples and in table 7 for healthy samples. Let's first note that even if there were only 5 healthy samples, some statistically significant correlation were found between DNAm acceleration and C2:CP genes sets expression. Out of the 4 gene sets whose expression was correlated to DNAm acceleration in healthy samples, 3 of them had also their expression correlated to DNAm acceleration in tumour samples. Therefore those correlations are not specific to tumour cells. Then it is also important to note that all the significant correlations where positive i.e. the lower the DNAm acceleration, the lower the expression of those genes set. It is interesting to note that the C2:CP genes set whose expression had the most significant correlation in tumour cells with DNAm acceleration is about cytokines which have been proposed as lung cancer biomarkers and therapeutic targets [13]. Moreover the second most significant one, contains the genes coding for extracellular matrix proteins that have been shown to protect small cell lung cancer cells against apoptosis [14].

The pairwise correlation between genes expression and DNAm acceleration were computed with the model from equation 6 and the empirical Bayes method. For healthy samples, no gene expression was statistically significantly correlated with DNAm acceleration i.e. $w_{i,\text{acc}}$ was not statistically significantly different from 0. For tumour samples, out the 17884 genes that remained after the filtering by the function filterByExpr, 1488 had their expression positively correlated with DNAm acceleration (i.e. $w_{i,\text{acc}}$ was statistically significantly greater than 0) and 1028 had their expression negatively correlated with DNAm acceleration (i.e. $w_{i,\text{acc}}$ was statistically significantly lower than 0). A volcano plot of the $w_{i,\text{acc}}$ for the tumour samples is given in figure 6.

A more in depth analysis could be performed by reviewing the literature on the genes and genes sets whose expression was the most statistically significantly correlated with DNAm acceleration but this was beyond the scope of this project.

C2:CP Pathway	Direction	BH FDR
SA_MMP_CYTOKINE_CONNECTION	Up	1.91×10^6
NABA_ECM_AFFILIATED	Up	1.24×10^5
NABA_MATRISOME	Up	1.24×10^5
NABA_MATRISOME_ASSOCIATED	Up	1.69×10^{-5}
NABA_CORE_MATRISOME	Up	0.00031
NABA_SECRETED_FACTORS	Up	0.000541
NABA_ECM_REGULATORS	Up	0.00057
NABA_PROTEOGLYCANS	Up	0.00057
SIG_PIP3_SIGNALING_IN_B_LYMPHOCYTES	Up	0.000893
NABA_ECM_GLYCOPROTEINS	Up	0.00232
SIG_BCR_SIGNALING_PATHWAY	Up	0.0046
NABA_COLLAGENS	Up	0.0196
SA_FAS_SIGNALING	Up	0.0471

Table 6: Significant correlations between DNAm acceleration in tumour cells with C2:CP gene expressions. If the direction is set to Up, it means that the higher the DNAm acceleration, the more expressed are those pathways. BH FDR is the Benjamini and Hochberg FDR adjusted p value.

C2:CP Pathway	Direction	BH FDR
NABA_CORE_MATRISOME	Up	0.0009
NABA_COLLAGENS	Up	0.0037
NABA_ECM_GLYCOPROTEINS	Up	0.0037
SA_REG_CASCADE_OF_CYCLIN_EXPR	Up	0.024

Table 7: Significant correlations between DNAm acceleration in healthy cells with C2:CP gene expressions. If the direction is set to Up, it means that the higher the DNAm acceleration, the more expressed are those pathways. BH FDR is the Benjamini and Hochberg FDR adjusted p value.

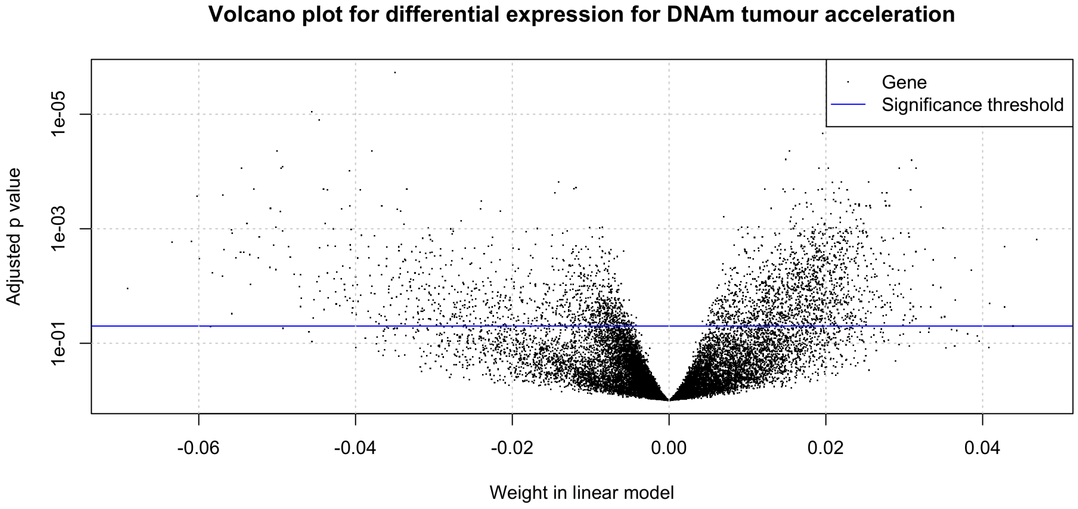

Figure 6: Volcano plot of the obtained weights for the linear model of the genes expressions with the DNAm tumour acceleration.

4 DNAm age versus number of somatic mutations and copy number variations in LUSC samples

The correlation between DNAm age and number of somatic mutations was analysed with the Mutation Packager Calls data from the TCGA database. It contains a list of somatic mutations for each sample. For each tumour samples, the somatic mutations were counted and a scatter plot of the number of somatic mutations against the DNAm acceleration is given in figure 7. The Spearman correlation coefficient is $r = -0.22$ with an estimated p value of 0.0096. It is interesting to note that this small correlation between DNAm acceleration in tumour cells was found significant for other cancer types in [4] but not in LUSC tumours. This may be due to a different data set used in their analysis.

The correlation between copy number variation (cnv) of genes and DNAm acceleration was computed using the genome_wide_snp_6_segmented_scna_minus_germline_cnv_hg19 TCGA file. It contains the segment mean variable which is the logarithm of the probes intensity ratio for the tumour DNA with respect to the germline. This means that a positive segment mean indicates (if large enough), that the corresponding genome sequence contains more copies than in the germline and a negative segment mean indicates (if large enough) fewer copies than in the germline. The classification problem of the segment mean value significance (i.e. if it is statistically significantly different from 0) is not a simple problem. One simple approach is to use an arbitrary threshold for the segment mean value to be considered significant. In this report it was decided to use the threshold = 0.2 (and -0.2 for negative values) as it was done in [15]. Then the genome sequences that had a segment mean value significantly different from 0 were mapped to the human genome hg19 with the R package Homo.sapiens, to identify which genes had an abnormal copy number. A scatter plot of the DNAm acceleration of tumour cells versus the number of genes with higher copy number that expected is given in figure 8. The same plot but for the number of genes with lower copy numbers than expected is given in figure 9. Both the number of genes with abnormally high and low copy numbers were correlated with the DNAm acceleration in the tumour samples : Spearman coefficient $r = -0.25$ with an estimated p value of 4.5×10^{-5} and Spearman coefficient $r = -0.26$ with an estimated p value of 3.1×10^{-5} respectively. This seemed to indicate that the correlation was independent on the direction (higher or lower) of the copy number difference. Indeed, for the correlation of the number of genes was abnormal copy numbers (both low and high), the Spearman coefficient was $r = -0.28$ with an estimated p value of 6.8×10^{-6}. This indicates a small negative correlation between DNAm acceleration and number of genes with abnormal copy numbers in tumour cells. A more in depth analysis could be performed to determine if the correlation is higher for specific gene sets such as the ones from the C2:CP dataset.

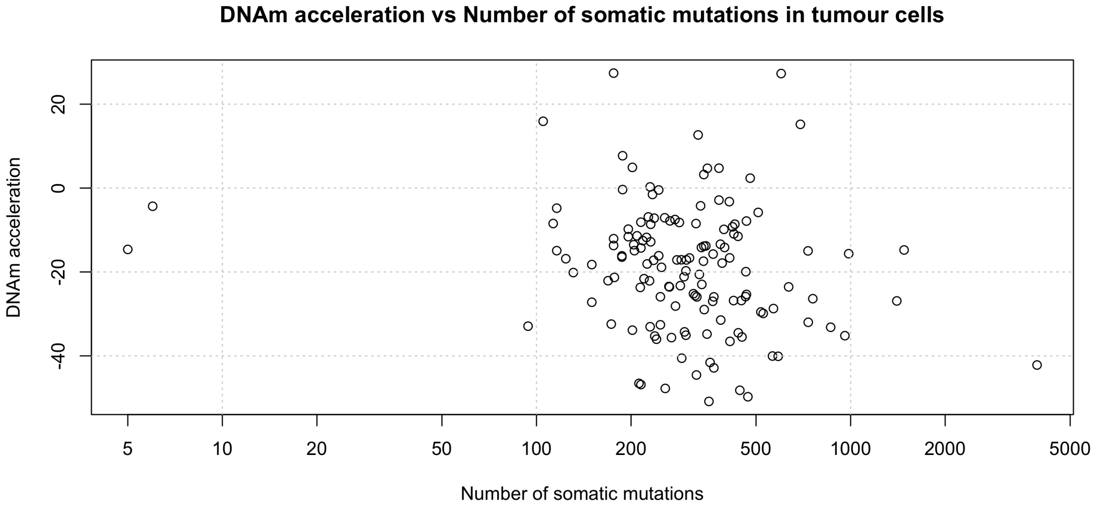

Figure 7: DNAm acceleration vs Number of somatic mutations in tumour cells. Each dot corresponds to a tumour sample.

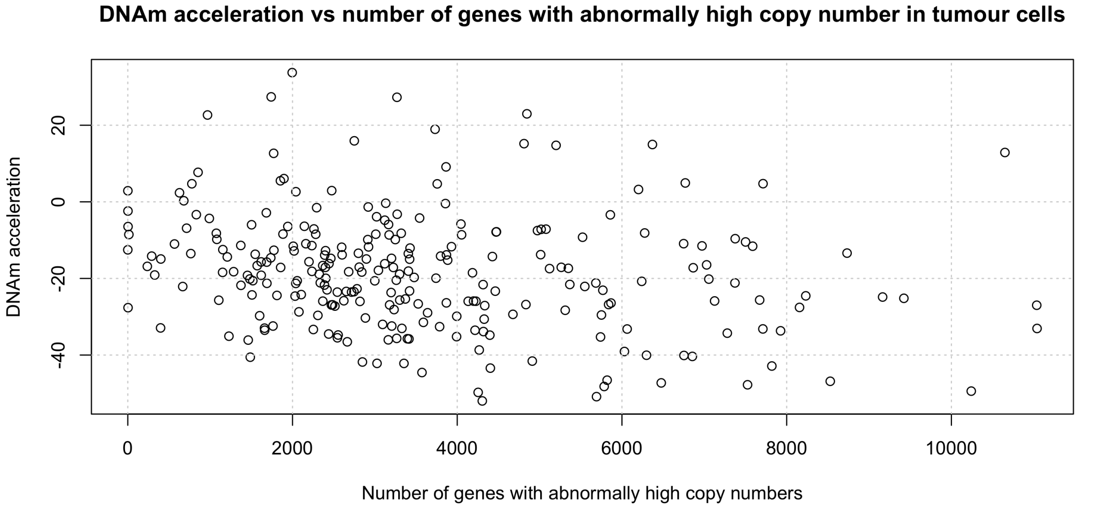

Figure 8: DNAm acceleration vs Number of genes with abnormally high copy numbers in tumour cells. Each dot corresponds to a tumour sample.

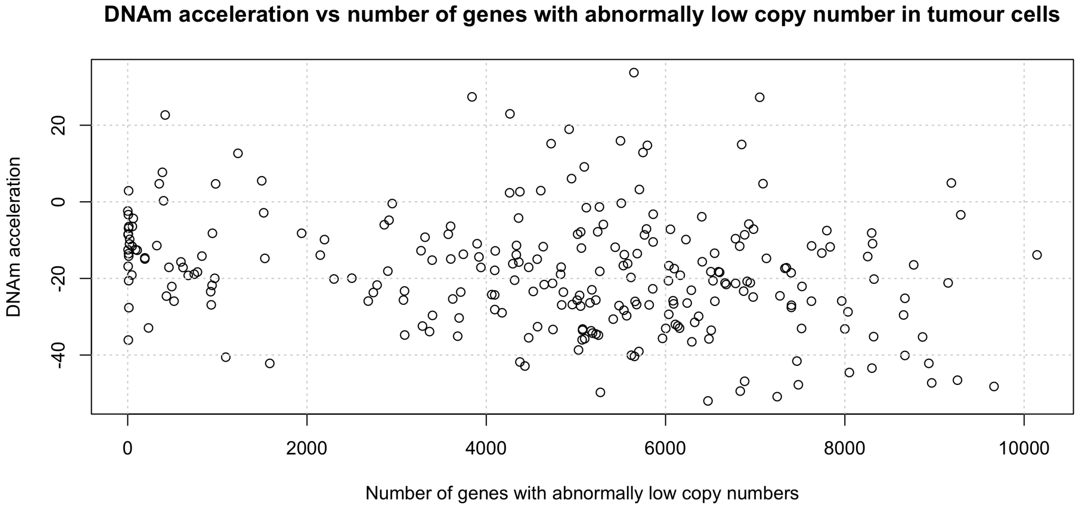

Figure 9: DNAm acceleration vs Number of genes with abnormally low numbers in tumour cells. Each dot corresponds to a tumour sample.

5 Conclusion

A statistically significant decrease in tumour DNA methylation age was observed in LUSC tumour cells. However no statistically significant correlations between this decrease and the patient's clinical variables were found. Due to the current cost decrease in sequencing technologies, maybe in the future more DNA methylation data of LUSC patient will be available and similar investigations to the one presented in this paper can be carried out to uncover unknown correlations.

Using RNAseq data from the tumours, it was found that some genes sets from the MSigDb C2:CP dataset had their expression correlated with the decrease in DNA methylation age. While three of them were also correlated to the decrease in DNAm age in healthy cells, the others were not. The two genes sets which had their gene expression correlated to DNAm acceleration with highest significance, code for protein families that were linked to lung cancer. The most significant gene set codes for the cytokines family who have been proposed as a biomarker for lung cancer and therapeutic targets [13]. The second most significant one, codes for extracellular matrix proteins that have been shown to protect small cell lung cancer cells against apoptosis [14].

Finally, a small, but statistically significant, negative correlation between the DNA methylation age and both the number of somatic mutations and the number of genes with abnormal copy number in tumour cells was observed.

While in this study some statistically significant correlations were found, a more in depth analysis should be performed to better understand them in order to extract biological and/or clinical meaningful information.

6 Appendix A: Normality and variance homogeneity conditions for ANOVA like tests

The ANOVA test requires the data to be normally distributed in each group and the variance of the data not to vary across groups. The homogeneity of the variance across groups is tested with Levene test (using the median instead of the mean for the centre in each group) and the normality is computed with the Shapiro-Wilk test on the residuals of the data (data minus group mean). The results of these two tests for the different clinical variables is given in table 8. The significance threshold used for those tests is $\alpha = 5\%$. It was decided not to use any correction for the significance threshold (such as Bonferonni or Sidak) since false negative results in those tests are much problematic that false positives. If the variance homogeneity is violated and not the normality, the Welch one way ANOVA test is performed instead of the ANOVA. If the normality is violated but not the variance homogeneity, the Kruskal-Wallis test is performed. If there is a false positive in the Shapiro-Wilk or Levene's test, using another test than ANOVA is still statistically correct, the only issue is that Kruskal-Wallis and Welch's ANOVA have less statistical power. On the other hand, a false negative result could lead to an erroneous usage of the ANOVA test which could invalidate the ANOVA results.

Variable	**DNAm tumour age**			**DNAm tumour age**		
	Shapiro p	**Levene** p	**Test**	**Shapiro** p	**Levene** p	**Test**
T stage		0.04	Welch's ANOVA	0.001	0.24	Kruskal-Wallis
N stage	0.02	0.52	Kruskal-Wallis	0.001	0.46	Kruskal-Wallis
M stage	0.003	0.07	Kruskal-Wallis	0.0001	0.08	Kruskal-Wallis
Pathologic stage	0.02	0.30	Kruskal-Wallis	0.0004	0.36	Kruskal-Wallis
Gender	0.04	0.71	Kruskal-Wallis	0.001	0.40	Kruskal-Wallis
Radiation therapy	0.11	0.86	ANOVA	0.02	0.97	Kruskal-Wallis
Histological type	0.02	0.75	Kruskal-Wallis	0.001	0.07	Kruskal-Wallis
Residual tumour	0.007	0.13	Kruskal-Wallis	0.002	0.39	Kruskal-Wallis
Ethnicity	0.49	0.35	ANOVA	0.21	0.27	ANOVA
Race	0.43	0.99	ANOVA	0.42	0.39	ANOVA

Table 8: Shapiro-Wilk and Levene test results used to determine which correlation test to perform.

References

[1] Mario F Fraga and Manel Esteller. Epigenetics and aging: the targets and the marks. *Trends in Genetics*, 23(8):413–418, 2007.

[2] Steve Horvath. Dna methylation age of human tissues and cell types. *Genome biology*, 14(10):3156, 2013.

[3] Sven Bocklandt, Wen Lin, Mary E Sehl, Francisco J Sánchez, Janet S Sinsheimer, Steve Horvath, and Eric Vilain. Epigenetic predictor of age. *PloS one*, 6(6), 2011.

[4] Steve Horvath. Erratum to: Dna methylation age of human tissues and cell types. *Genome biology*, 16(1):96, 2015.

[5] Alfredo Morabia and Ernst L Wynder. Cigarette smoking and lung cancer cell types. *Cancer*, 68(9):2074–2078, 1991.

[6] DJ Best and DE Roberts. Algorithm as 89: the upper tail probabilities of spearman's rho. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(3):377–379, 1975.

[7] Selena Hsin Feng and Su-Tso Yang. The new 8th tnm staging system of lung cancer and its potential imaging interpretation pitfalls and limitations with ct image demonstrations. *Diagnostic and Interventional Radiology*, 25(4):270, 2019.

[8] Paul Hermanek and Christian Wittekind. Residual tumor (r) classification and prognosis. In *Seminars in surgical oncology*, volume 10, pages 12–20. Wiley Online Library, 1994.

[9] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12(1):323, 2011.

[10] Di Wu and Gordon K Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research*, 40(17):e133–e133, 2012.

[11] Yunshun Chen, Davis McCarthy, Mark Robinson, and Gordon K Smyth. edger: differential expression analysis of digital gene expression data user's guide. *Bioconductor User's Guide. Available online: http://www. bioconductor. org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide. pdf (accessed on 28 May 2020)*, 2014.

[12] Mark D Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of rna-seq data. *Genome biology*, 11(3):R25, 2010.

[13] Ángela Marrugal, Laura Ojeda, Luis Paz-Ares, Sonia Molina-Pinelo, and Irene Ferrer. Proteomic-based approaches for the study of cytokines in lung cancer. *Disease markers*, 2016, 2016.

[14] Tariq Sethi, Robert C Rintoul, Sarah M Moore, Alison C MacKinnon, Donald Salter, Chin Choo, Edwin R Chilvers, Ian Dransfield, Seamas C Donnelly, Robert Strieter, et al. Extracellular matrix proteins protect small cell lung cancer cells against apoptosis: a mechanism for small cell lung cancer growth and drug resistance in vivo. *Nature medicine*, 5(6):662–668, 1999.

[15] Saurabh V Laddha, Shridar Ganesan, Chang S Chan, and Eileen White. Mutational landscape of the essential autophagy gene becn1 in human cancers. *Molecular cancer research*, 12(4):485–490, 2014.