BINF-F401 : Computational Methods for Functional Genomics

DNA methylation age in lung adenocarcinoma cancer.

Chloé Terwagne
registration number: 000409683

August 2020

Introduction

This report presents the results of the study of the lung adenocarcinoma cancer associated with the estimated DNA methylation age using the publicly available data from The Cancer Genome Atlas (TCGA) project.

The methylation of DNA is a biological process, catalyzed by the DNA methyltransferases, by which methyl groups are added to the DNA bases (cytosine and adenine). The DNA methylation plays a critical role in epigenetic silencing of transcription[1]. DNA methylation is a key process for normal development, notably playing a role in X-chromosome inactivation, repression of transposable elements, aging, and carcinogenesis. The DNAm clock, proposed in 2013 by Steve Horvath [2] (and corrected in 2015[3]), is a highly accurate molecular biomarker of aging which aims to estimate the biological age based on DNA methylation levels (abbreviated by DNAm).

Lung cancer has a higher mortality rate than any other cancer for both, men and women[4]. Lung cancer accounts for about 28 percent of all cancer deaths. Along with the lung adenocarcinoma cancer (abbreviated by LUAD in this report), the lung squamous cell carcinoma are the two subtypes of the non-small cell lung cancer. LUAD is the most common form of lung cancer accounting for about 30 percent of all lung cancers. The lung adenocarcinoma cancer tends to grow slower than all the other cancers[5]. LUAD cancer has a high rate of somatic mutation throughout the genome as well as genomic rearrangements[6]. Finally, tobacco is the leading cause in LUAD and only a minority (10–15 percent) of people diagnosed are never-smokers [7].

This work is based on published data from The Cancer Genome Atlas (TCGA) project. The TCGA project launched in 2005 under the supervision of the National Cancer Institute's Center for Cancer Genomics and the National Human Genome Research Institute aims at characterising different types of human cancer. Currently, 33 cancer types are molecularly characterised using, for instance, genome sequencing. Over the past years, large amounts of data generated by the TCGA program (including genomic, transcriptomic, epigenomic and proteomic data) have allowed to improve both treatments and diagnosis of cancers.

The files used for this project are download via Firehose (http://gdac.broadinstitute.org/) using TCGA data version 01.28.2016

- LUAD cancer: 'Clinical_pick_Tier1', Patients clinical annotation (including chronological age, demographic information, treatment information, survival data, etc).
- LUAD cancer: 'illuminahiseq_rnaseqv2_RSEM_genes_normalized', mRNA genes expression estimated from TCGA mRNA-seq data.
- LUAD cancer: 'Mutation_Packager_Calls', Mutation Annotation Format (MAF).
- LUAD cancer: 'genome_wide_snp_6_segmented_scna_minus_germline_cnv_hg19', file containing the copy number variation (CNV).

The last file, LUAD-7.Rda, contains the estimated DNAm age computed by Vincent Detours.

To facilitate the reproducibility; the integrality of my code and a text format (sessionInfo.txt) are available on GitHub repository (link : https://github.com/Chloe-Terwagne/LUAD.functional-genomics). These ones contain version information about R, the OS and attached or loaded packages,

Through this project, several relationships between tissues, normal and cancer ones, and the DNAm acceleration will be presented. Firstly, the correlation between the chronological age and the DNAm age is described. Secondly, relationships between diverse patients' clinical variables and their corresponding DNAm age in cancer tissues. Thirdly, the one between DNAm age in cancers is used in assessing genes and pathways abnormally over- or under-expressed. Finally, the correlation between DNAm age and the number of somatic mutations and the relation between DNAm age and the number of DNA copy number breakpoints.

1 Correlation between chronological and DNAm age.

This section presents the relation between the chronological age (obtained from the patients' clinical annotations of the TCGA, 'ClinicalpickTier1') and the estimated DNA methylation age (precomputed by Vincent Detours, 'LUAD-7.Rda') for the lung adenocarcinoma.

After merging the patients' clinical annotations (522 observations) and the DNAm_age (410 samples) by patient's number[1], 315 observations were obtained[2] (56 healthy samples and 259 samples from cancer tissues).

The relation is measure by the spearman's correlation. The spearman coefficient is a non-parametric measure of rank correlation. This coefficient assesses how two variables are related using a monotonic function. The spearman's test used for all spearman's correlation in this project is the cor.test function in R. The p-values of the cor.test function are computed using algorithm AS 89[3]. The cor.test function can only calculate the exact p-values when there are no ties. However, the chronological age of the patients contains multiples ties as shown for tumor sample data in table 1. Therefore, finding a more appropriate calculation method for the data could improve the exactness of the p-value.

Age	38	41	42	43	45	47	48	49	50	51	52	53	54	55	56	57	
Frequency	1	1	5	2	2	2	1	1	3	6	4	4	3	4	3	4	
Age	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73	
Frequency	9	14	13	13	8	5	7	10	5	10	7	7	8	9	13	7	
Age	74	75	76	77	78	79	80	81	82	83	85	86					
Frequency	12	10	8	3	1	5	2	2	3	1	2	1					

Table 1: Chronological age and their corresponding frequency for the tumor dataset.

1.1 Results.

1.1.1 Correlation between DNAm age for cancer tissues and chronological age.

The figure 1 is a plot of the 241 observations (18 out of the 259 tumor samples have no value for the chronological age.) of the DNAm age for cancer samples[4] versus the chronological age of the patients. The spearman's correlation obtained is weakly positive (rho = 0.2089718) and

[1]The patient/participant number is extracted from the TCGA barcode (Additional TCGA barcode informations are available at https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/).

[2]20 out of the 315 observations have no value for the chronological age (years_to_birth) in their patient clinical information, thus these 20 observations are not taking into account by the spearman's correlation (cor.test function in R).

[3]The AS 89 algorithm has two methods, it uses the "exact" one if $n < 9$ and the "semi-exact" one if $9 < n < 1290$. In our case, the p-value is thus always calculated by the latter method.[8]

[4]Sample type between 01-09 indicating that the sample is the type "tumor" in TCGA barcode names.

significant with a p-value < 0.05 (p-value $= 0.001101$). The patients' age varies between 38 and 86 years while the DNAm age varies more widely between 1.49 and 131.67.

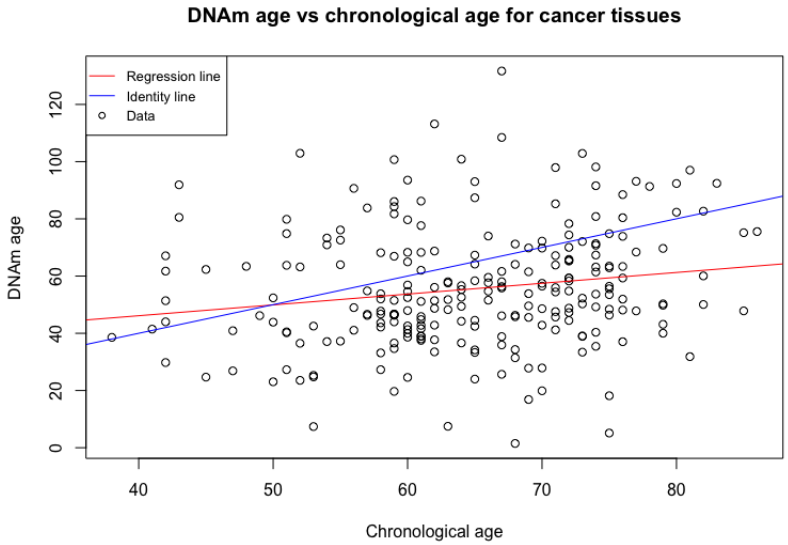

Figure 1: Scatter plot of DNAm age (in years) versus chronological age (in years) for cancer tissues. The identity line (x = y), the regression line and the data are represented in blue, in red and by empty black points; respectively.

1.1.2 Correlation between DNAm age for healthy tissues and chronological age.

The 54 data obtained for normal samples[5] (2 out of the 56 healthy samples have no value for the chronological age.) are shown in figure 2. As expected, the spearman correlation is strongly positive with a rho = 0.7568633 and highly significant (with a p-value of 3.567e-11). A large majority of the data have a chronological age greater than the DNAm age.

The linear model for these data is define by the equation below:

$$A_{DNAmN} = 10.7661 + 0.7917 A_{chronological} \tag{1}$$

Where A_{DNAmN} is the estimated DNAm age of normal samples and $A_{chronological}$ is the chronological age. This linear model is obtained using an ordinary least squares method. The intercept and slope estimation, standard error and p-value are in table 2.

	Estimation	Standard error	P-value
Intercept	10.7661	5.4866	0.0551
Slope	0.7017	0.0835	2.95e-11

Table 2: Summary of the linear model of DNAm versus chronological age for normal tissues.

[5]Sample type between 10-19 indicating that the sample is the type "normal" in the TCGA barcode names.

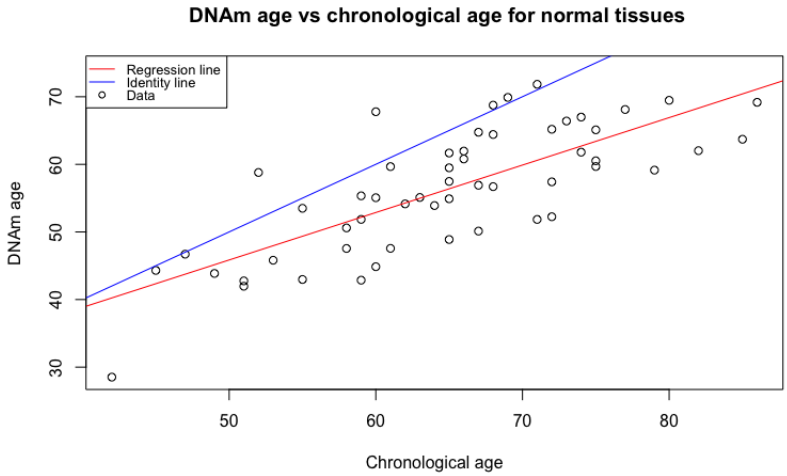

Figure 2: Scatter plot of DNAm (in years) versus chronological age (in years) for normal tissues. The identity line (x = y), the regression line and the data are represented in blue, in red and by empty black points; respectively.

1.1.3 Correlation between DNAm age of tumors and their patient matched normal tissue.

58 data are obtained to compare DNAm age of tumors and their patient-matched normal tissue. Here, 4 more observations are generated than in figure 2, this is explained by the facts that some patients have more than one tumor sample. Thus, some samples from normal tissues are duplicated to enable the comparison with two tumor samples from different plates[6]. The patients' numbers concerned are 4112, 3918, 2668, 2665 and 2656. The spearman coefficient measuring the correlation between DNAm age of tumor samples and their patient-matched normal sample is not significant with a p-value of 0.5753.

1.1.4 Correlation between DNAm acceleration for cancer tissues and chronological age.

The acceleration[7] is defined as (the acceleration choice is justified in the discussion section):

$$a_T = A_{DNAmT}/A_{DNAmN} \tag{2}$$

Where a_T is the tumor age acceleration, A_{DNAmT} is the DNA methylation age for cancer tissues from the data and A_{DNAmN} is the estimated DNA methylation age predict thanks to the equation 1.

When the tumor age acceleration equal to one, there is no acceleration (meaning that estimated DNA methylation age from normal samples equals DNA methylation age for cancer tissues). When the tumor age acceleration is greater (lower) than one, DNAm age of cancer tissues is higher(lower) than expected.

The p-value for the spearman's correlation between the tumor age acceleration and the chronological age is 0.1153. Thus, there is no statistically significant correlation as shown in figure 3.

[6]The plate's position of the sample in a sequence of 96-well plates is extracted from the TCGA barcode name.

[7]This is not technically an acceleration, the 'average acceleration' is defined in Steve Horvath's article[2] by the average difference between the DNAm age and the chronological age. This acceleration enables the comparison between tissues and organs. Thus, it's possible to ascertain if a given tissue has a lower or higher DNAm age than expected.

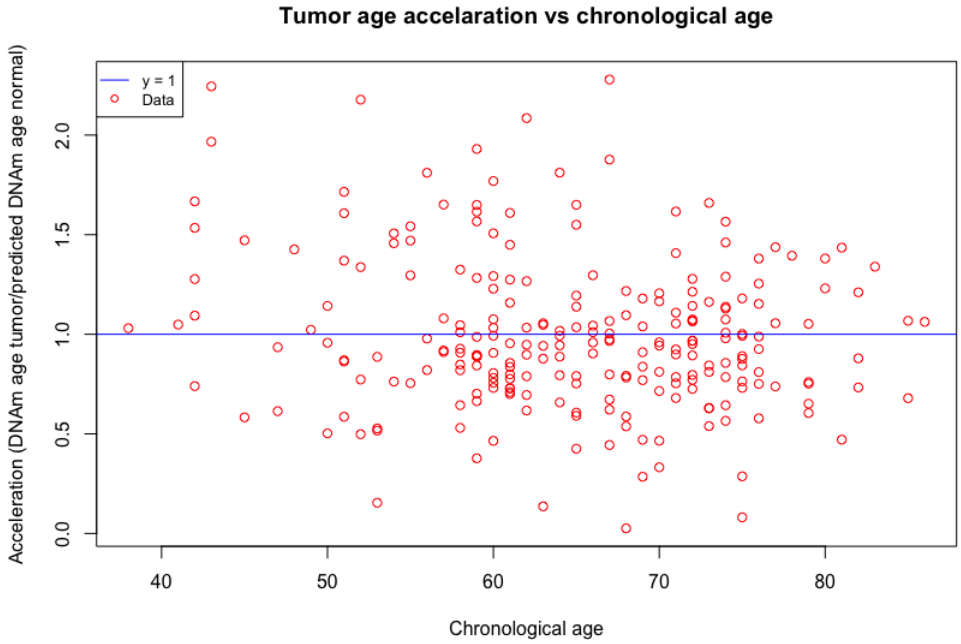

Figure 3: Scatter plot of tumor age acceleration versus chronological age (in years). The horizontal line for y = 1 and the data are represented in blue and by empty red points; respectively.

1.2 Discussion.

As expected, in normal tissues the DNAm age is highly correlated with the chronological age as shown in figure 2. Indeed, the DNAm age is calculated to be non-biased [2]. However, the DNAm age is mostly smaller than the chronological age. Two possibilities justify this observation. The first one is that the DNAm age was calculated by an alternative method than the one described in Horvath's paper. The second possibility is the lung is an organ (and/or the tissues that compound it) ageing at the different speed that the average speed given that some inter-organ variation exists. In this case, the lung's biological age (DNAm age) evolve slower than the chronological age.

In figure 1, cancer tissues have a very variable DNAm age (greater or smaller than chronological age) and they are weakly positively correlated with the chronological age of patients. It's evident that the biological age is not as calibrated in the cancer samples as it's the case for healthy samples in figure 2.

There is no correlation between acceleration and chronological age as illustrated in figure 3. To measure the tumor age acceleration several options exist. A first one is to divide (or subtract) the DNAm age of tumor by DNAm age normal but this option use only a small part of the data (58 observations). A second option is to divide (or subtract) DNAm age of tumor by the chronological age of the patient. This option is not ideal because DNAm age of normal tissues doesn't equal the chronological age as shown in equation 1, the slope is significantly different of 1 and the intercept of 0. A better option is to use an ordinary least squares linear regression from figure 2 to predict the DNAm age normal for each tumor sample. Thus, the acceleration is defined by the division of the DNAm age of tumor by DNAm age normal predicted in equation 1. This method avoids excluding a large part of the data set as it's the case in the first option. This last option is chosen for this report.

2 Correlation between clinical variables and DNAm age in cancers.

In this section, the correlations between the DNAm age in cancer tissues and the diverse clinical variables is studied. In the first and second subsection, the continuous and categorical variables are analysed, respectively.

To counteract the problem of multiple comparisons, the šidák method is applied to compute the adjusted p-value. This step prevent the inflation of false positive rate while doing multiple comparisons. The šidák correction has more power than the Bonferonni correction and assumes that each comparison is independent of the others (Which is not always the case, as TNM stage and the pathologic stage or for the smoking period and time period etc.). Šidák is slightly less conservative than the Bonferonni method. The šidák correction formula is:

$$\alpha_{SIDAK} = 1 - (1 - \alpha)^{1/m} \tag{3}$$

In this section, to study the DNAm acceleration 14 tests (10 for categorical and 4 for continuous variables) are applied. Thus, α_{SIDAK} at level $\alpha = 5\%$ is 0.003657 (rounded to 6 decimal places).

In this section, all indications concerning the number of samples for each variable or variable's category is based on the number of samples used to correlate the given variable with the DNAm age acceleration. The number of samples used for the DNAm age is a bit higher. This is due to the missing value for chronological age for some clinical data resulting in a slight decrease of DNAm age acceleration data set. Although this report contains the results for the correlations with the DNAm age and with DNAm age acceleration for cancer tissues, only the acceleration is relevant and will be analysed in details.

2.1 Continuous variables.

In table 3, the different clinical variables studied are describe. The number packs years smoked and the karnofsky performance score are directly from the clinical annotation files. However, the smoking period and total number packs smoked are calculated as defined in table 3's description. The smoking period assumes that the patient has not stopped smoking until his diagnosis.

Variable (number of samples)	**Description**
smoking_period (133)	period (in years) of smoking from the year_of_tobacco_smoking_onset to the date_of_initial_pathologic_diagnosis
number_packs_years_smoked (163)	
total_number_packs_smoked (125)	number_packs_years_smoked multiply by the smoking_period
karnofsky_performance_score (65)	An index designed for classifying the functional impairment of a patient. It measures the ability of patients to perform ordinary tasks.

Table 3: Description of the continuous variables.

The correlations for continuous variables is computed with the spearman coefficient (cor.test function in R). The results are presented in table 4, none of the results has p-value below the adjusted p-value which is 0.003657. Therefore there is no significant correlation for these four variables.

	Test	**DNAm age**		**Acceleration**	
		p-value	rho	p-value	rho
Smoking period	spearman	0.4507	0.06496079	0.1551	-0.1239688
Number packs years smoked	spearman	0.2126	-0.0952424	0.179	-0.1057835
Total number packs smoked	spearman	0.1105	-0.1434521	0.03801	-0.1858198
Karnofsky performance score	spearman	0.8981	-0.01607638	0.7377	-0.04234449

Table 4: Results of the spearman correlations for the continuous variables. The adjusted p-value is 0.003657.

2.2 Categorical variables.

In table 5, the different categorical variables are presented. Some of them were regrouped: the M staging group m1 (9 observations) and their subgroup m1a (1 observation) and m1b (1 observation) [8] are reorganised into one group of 11 observations. The reorganisation is proceeded with t1a and t1b reorganise under t1 and with t2a and t2b reorganise under t2. This reorganisation is done to have more homogeneous groups since the other categories don't have subgroup.

Variable (number of samples)	**Categories** (number of observation for each category)	**Description**	**New categories**
pathologic_stage (237)	stage ia (60) stage ib (67) stage iia (23) stage iib (32) stage iiia (39) stage iiib (5) stage iv (11)	The pathologic stage of the cancer describe the combination of the T, N, and M classifications. stage ia: t1a—n0—M0 or t1b—n0—M0 stage ib: t2a—n0—M0, stage iia: t2b—n0—m0 stage iib: t1—n1—m0 or t2—n1—m0 or t3—n0—m0 stage iiia: t1—n2—m0 or t2—n2—m0 or t3—n1—m0 or t4—n0/n1—m0 stage iiib: t1/t2—n3—m0 or t3—n2—m0 or t4—n2—m0 stage iv: any t— any n— m1	
pathology_T_stage (241)	t1 (34), t1a (22), t1b (20) t2 (84), t2a (38), t2b (10) t3 (23) t4 (9) tx (1)	Describes the size and/or extension of the primary tumor. t1 ≤3cm (t1a : minimal invasive adenocarcinoma, 1<t1b≤2cm); 3<t2≤5cm (3<t2a≤4, 4<t2b≤5); 5<t3≤7cm; t4>7cm; tx: The primary tumor cannot be evaluated.	t1 (76) t2 (132) t3 (23) t4 (9) tx (1)
pathology_N_stage (240)	n0 (151) n1 (43) n2 (40) nx (7)	Describes whether or not the cancer has reached nearby lymph nodes (N). n0: no regional lymph nodes metastasis n1: ipsilateral peribronchial and/or hilar and intrapulmonary nodes n2: ipsolateral mediastinal and/or subcarinal nodes nx: lymph nodes cannot be evaluated.	
pathology_M_stage (238)	m0 (170) m1 (9), m1a (1), m1b (1) mx (57)	Describes the presence of metastasis (M). A standard way of measuring spread of cancer to other parts of the body. m0: no distant metastasis m1: distant metastasis mx: cannot be evaluated for distant metastasis	m0 (170) m1 (11) mx (57)
gender (241)	female (126) male (115)		
radiation_therapy (217)	yes (35) no (182)		
histological_type (241)	lung acinar adenocarcinoma (5) lung adenocarcinoma-nos[9](154) lung bronchioloalveolar carcinoma nonmucinous (10) lung micropapillary adenocarcinoma (2) lung papillary adenocarcinoma (10) mucinous (colloid) carcinoma (4) lung adenocarcinoma mixed subtype (55) lung bronchioloalveolar carcinoma mucinous (1) lung mucinous adenocarcinoma (2) lung solid pattern predominant adenocarcinoma (1)		
residual_tumor (170)	r0 (151) r1 (5) r2 (3) rx (11)	r0:no residual tumor r1: microscopic residual tumor r2: macroscopic residual tumor rx: presence of residual tumor cannot be assessed	
race (221)	american indian or alaska native (1) white (190) asian (4) black or afican american (26)		
ethnicity (195)	hispanic or latino (5) not hispanic or latino (190)		

Table 5: Description of the categorical variables.

The categorical variables are studied using ANOVA test when equal variance assumption and normal distribution of residuals assumption aren't violated. Otherwise, alternatives to ANOVA test are used. If the normal distribution of residuals assumption is violated, Kruskal-Wallis is used. If only the equal variance assumption is violated, Welch ANOVA is used.

[8] m1a: regional metastatic disease defined as malignant pleural or pericardial nodules, as well as contralateral or bilateral pulmonary nodules. m1b: solitary extrathoracic metastasis

To test the variances homoskedasticity assumption, the Levene test is chosen. When the p-value is below 0.05: the null hypothesis of equal variances is rejected. So these variables violate the homogeneity of variance assumption needed for an ANOVA. Another option is to use the Bartlett test to compare the variances between several groups. However, Bartlett's test requires the data to be normally distributed. This is why Levene test is used. The Levene test is a more robust alternative to the Bartlett test, it is less sensitive to deviations from normality. Also, Bartlett test requires at least two observation for each category which is not the case for all our variable (including race or histological type).

To test the normal distribution of residuals assumption, the Shapiro test is used. The null hypothesis is that the variable follows a normal distribution. If the p-value is below 0.05 (at level 5 percent), H0 is rejected. The conditions of application of normal distribution are therefore not respected from the point of view of the normality of the residues.

A summary of the results obtained is shown in table 6 with p-values for the Shapiro and Levene tests along with the test to use depending if the ANOVA required assumptions are violated or not.

	DNAm age			**Acceleration**		
	Shapiro p-value	Levene p-value	Test	Shapiro p-value	Levene p-value	Test
Pathologic stage	0.001682	0.3297	Kruskal-Wallis	8.397e-06	0.1204	Kruskal-Wallis
Pathology T stage	0.02553	0.5968	Kruskal-Wallis	4.937e-05	0.3398	Kruskal-Wallis
Pathology N stage	0.006716	0.06468	Kruskal-Wallis	1.506e-05	0.2172	Kruskal-Wallis
Pathology M stage	0.005325	0.5262	Kruskal-Wallis	9.065e-06	0.287	Kruskal-Wallis
Gender	0.006633	0.5728	Kruskal-Wallis	2.114e-05	0.9872	Kruskal-Wallis
Radiation therapy	0.01539	0.7901	Kruskal-Wallis	0.0001015	0.5853	Kruskal-Wallis
Histological type	0.003425	0.2395	Kruskal-Wallis	1.162e-05	0.4561	Kruskal-Wallis
Residual tumor	0.05146	0.3256	ANOVA	0.0009773	0.4552	Kruskal-Wallis
Race	0.02465	0.1347	Kruskal-Wallis	6.018e-05	0.06843	Kruskal-Wallis
Ethnicity	0.01468	0.5764	Kruskal-Wallis	0.0004002	0.01556	Kruskal-Wallis

Table 6: Results of p-values for the Shapiro and Levene tests along with the test to use depending if the ANOVA required assumptions are violated or not at level 5 percent.

In table 7, the test used for each categorical variables is presented as well as the p-value relative to test obtained. Since the adjusted p-value at level 5 percent is 0.003657 none of these results are significant neither for DNAm age and DNAm age acceleration.

	DNAm age		**Acceleration**	
	Test	P-value	Test	P-value
Pathologic stage	Kruskal-Wallis	0.2634	Kruskal-Wallis	0.1411
Pathology T stage	Kruskal-Wallis	0.8223	Kruskal-Wallis	0.7658
Pathology N stage	Kruskal-Wallis	0.7192	Kruskal-Wallis	0.5079
Pathology M stage	Kruskal-Wallis	0.2275	Kruskal-Wallis	0.1632
Gender	Kruskal-Wallis	0.4724	Kruskal-Wallis	0.9351
radiation therapy	Kruskal-Wallis	0.8146	Kruskal-Wallis	0.8393
Histological type	Kruskal-Wallis	0.8375	Kruskal-Wallis	0.7622
Residual tumor	ANOVA	0.85	Kruskal-Wallis	0.8563
Race	Kruskal-Wallis	0.5312	Kruskal-Wallis	0.7927
Ethnicity	Kruskal-Wallis	0.4886	Kruskal-Wallis	0.3948

Table 7: Results of the tests performed on the categorical variables for the DNAm age and DNAm age acceleration. The adjusted p-value is 0.003657.

2.3 Survival analysis.

For the survival analysis, two different ways to define the right censoring data are possible. The first one is using "days to last known alive", the other one is to use the "days to last followup". "Days to last followup" is a better choice since the number of samples available is larger, 152 samples against 10 for "days to last known alive data".

The description of each variable used for the survival analysis is given in table 8.

Variable (number of samples)	**Description**
vital_status (241)	alive is vital status equal to 0 (152) and dead is vital status equal to 1 (89)
days_to_death (89)	Number of days to death (overall survival time)
days_to_last_followup (152)	Number of days to the last follow-up (NA if the vital status is 1)

Table 8: Variables used for the survival analysis.

As a first step, Kaplan-Meier curves are plot providing a first analyse. Kaplan-Meier survival curves plot probability of survival against time providing easily measure as median survival time. The Kaplan-Meier estimator is a non-parametric statistic used to estimate the survival function. This estimator takes into account right censoring data[9], and it's defined by:

$$\widehat{S}(t) = \prod_{i:\ t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{4}$$

Where:

$S(t)$ is the estimator of the survival function.

t_i is a time when at least one event happened.

d_i is the number of events that happened at time t_i.

n_i is the individuals known to have survived (have not yet had a event or been censored) up to time t_i.

To realise the Kaplan-Meier curves, the DNAm age acceleration subsetted into four quartiles categories (because Kaplan-Meier estimator works better with categorical variable than continuous variable). Then the four quartiles survival curves are plot (using survfit and ggsurvplot R functions) as shown in figure 4. Figure 4 shows no clear difference between survival times between the DNAm age acceleration quartiles which is statistically confirmed by the log-rank test p-value of 0.33.

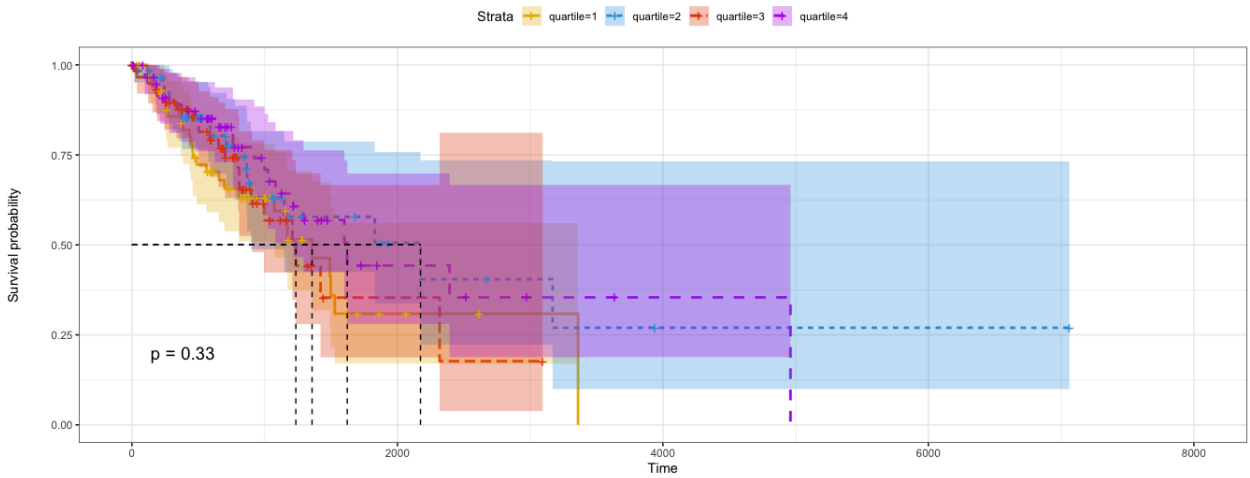

Figure 4: Kaplan-Meier curves of the survival analysis for the DNAm age acceleration quartiles with a log-rank test p-value of 0.33.

To analyse quantitative variable the Cox proportional hazards model is a better alternative to Kaplan-Meier curves and log-rank test. The Cox model is expressed by the hazard function. The hazard function defines the risk of dying at time t and is describe by:

$$h(t) = h_0(t)e^{(b_1x_1+b_2x_2+b_3x_3+...+b_px_p)} \tag{5}$$

Where:

$h(t)$ is the hazard function.

p is the number of covariates $(x_1, x_2, x_3, ..., x_p)$.

h_0 is the baseline hazard.

$b_1, b_2, ..., b_p$ are the coefficients measuring the impact of the covariates.

The results of the Cox proportional hazards model analysis are displayed in table 9. The analysis is performed with only one covariate (p =1), the DNAm age acceleration, giving the following equation for the hazard function : $h(t) = h_0(t)e^{bx}$ where x is the DNAm age acceleration and b the coefficient. From the table of results below, the coefficient b of the DNAm age acceleration is not statistically significantly different from 0 at level 5 percent.

The results for the DNAm age presented in table 9 are also not significant, they are given as requested by the assignment but working with the DNAm age is not as pertinent as working with the DNAm age acceleration.

	DNAm age		**Acceleration**	
	b	P-value	b	P-value
cox analysis	-0.007094	0.164	-0.3047	0.306

Table 9: Results of the Cox proportional hazards model for DNAm age and DNAm age acceleration.

3 Genes/pathways having expression associated with DNAm age acceleration in LUAD cancer.

The integrity of the code of this section is available in GitHub (link:https://github.com/Chloe-Terwagne/LUAD.functional-genomics) in the code folder genomics_section_3_4.

The "sample type" variable is created to define whether it's a healthy or tumor samples and the "tumor type" variable explicit which tumor it is (2 different kind of tumor is present in the dataset, 01 and 02). This why the files merging was done not only by "patient number" and "sample type" but also by the "tumor type" variable. After preprocessing the data, two mRNA-seq datasets are obtained. One for the tumor samples containing 20531 genes expression for 240 samples and the normal (healthy) dataset containing 20531 genes expression for 21 samples. In this section, the decision to also study the healthy samples is motivated by the possibility to compare the results of both normal and tumor samples and enable to define if a high expressed gene is specific or not to the tumor samples.

The major steps before performing the differential expression analysis are:

- The creation of a design matrix with the clinical information and DGE objects with the mRNA-seq datasets for both tumor and healthy samples
- Filtering out genes with a low amount of counts using the R function filterByExpr. After the filtering, the tumor and healthy datasets contains 17649 and 16552 genes, respectively. Removing the low-expressed genes is necessary to avoid the noise of genes with low counts. Moreover, filtering improves the reliability of the mean-variance relationship in the data estimation[10].

- The normalisation of the data is performed by the R function calcNormFactors using the default method 'TMM', trimmed mean of M-values. The normalisation helps to minimise technical variation and batch effects.
- The data are transformed by the voom R function. This converts the counts data to log2-counts per million (logCPM). Then, voom function estimates the mean-variance relationship to obtain weights for each gene and each sample[11].

3.1 Correlation between the C2:CP gene expressions and DNAm age acceleration.

To perform the pathway analysis, the Homo sapiens C2:CP (curated curated gene sets:canonical pathways, 3696 gene sets) is loaded from the msigdb database using the R function msigdbr. Then the R function camera is ran. The name camera stands for Competitive Gene Set Test Accounting For Inter-Gene Correlation and define if a set of genes is highly ranked relative to other genes in terms of differential expression, taking into account inter-gene correlation[12].

3.1.1 Results.

The statistically significant results generated for normal and tumor samples are shown is tables 11 and 10, respectively. All the significant results are positively correlated and set to the direction "Up" meaning that the higher the DNAm age acceleration is, the more the gene set is expressed. The NABA SECRETED FACTORS gene set expression is correlated with the DNAm age acceleration for both healthy and cancer samples. Thus, this correlation is not specific to tumor samples.

C2:CP ID	C2:CP name	Direction	Adjusted p-value
M11736	SA MMP CYTOKINE CONNECTION	Up	0.0008456361
M1315	SIG PIP3 SIGNALING IN B LYMPHOCYTES	Up	0.0033125640
M5880	NABA ECM AFFILIATED	Up	0.0158829677
M5885	NABA MATRISOME ASSOCIATED	Up	0.0250246944
M5882	NABA PROTEOGLYCANS	Up	0.0368790970
M8626	SIG BCR SIGNALING PATHWAY	Up	0.0368790970
M5883	NABA SECRETED FACTORS	Up	0.0368790970
M5060	SA FAS SIGNALING	Up	0.0368790970
M5889	NABA MATRISOME	Up	0.0368790970

Table 10: Significant results of the correlation between the C2:CP gene expressions and the DNAm age acceleration of 240 tumor samples. The adjusted p-value is obtained with the Benjamini-Hochberg false positive rate method.

The most significant result in table 10 is relate to cytokines. Indeed, cytokines play a key role as they modulate the immune and inflammatory response in non-small cell lung cancer[13]. Furthermore, the concentrations of circulating cytokines in serum are associated with lung cancer survival[14]. Moreover, other studies suggest that the immune microenvironment surrounding the tumor is influenced by interactions. These interactions are enabled to shift the environment to an anti or pro-tumor environment in non-small cell lung cancer. These interactions include tumor, immune cells (including lymphocytes B, which is the second most significant result) and cytokines[15] [16].

C2:CP ID	C2:CP name	Direction	Adjusted p-value
M5887	NABA BASEMENT MEMBRANES	Up	0.0002544797
M1718	SIG IL4RECEPTOR IN B LYPHOCYTES	Up	0.0013254837
M7955	SIG INSULIN RECEPTOR PATHWAY IN CARDIAC MYOCYTES	Up	0.0013254837
M295	SIG PIP3 SIGNALING IN CARDIAC MYOCTES	Up	0.0017933960
M5493	WNT SIGNALING	Up	0.0018850496
M5884	NABA CORE MATRISOME	Up	0.0064850531
M3008	NABA ECM GLYCOPROTEINS	Up	0.0064850531
M5883	NABA SECRETED FACTORS	Up	0.0281142465

Table 11: Significant results of the correlation between the C2:CP gene expressions and the DNAm age acceleration of 21 normal samples. The adjusted p-value is obtained with the Benjamini-Hochberg false positive rate method.

3.2 DNAm age acceleration and gene expressions.

The pairwise correlation between the DNAm age acceleration and genes expression is realised with the limma package using lmFit and constrast.fit R functions. These functions can be used either with microarray or mRNAseq data. lmFit fit a series of pairwise linear models between the DNAm age acceleration and the value of the expression for each gene. Then the R function eBayes compute several statistics by empirical Bayes moderation. These statistics can be used to rank the differentially expressed genes.

3.2.1 Results.

Figure 5 is a table generated using topTable R function ranking top genes sorted by p-value.

	logFC	AveExpr	t	P.Value	adj.P.Val	B
644168	3.495281	-4.214724	9.481339	2.476485e-18	4.370749e-14	29.60251
23532	-6.131752	1.069706	-9.237411	1.347349e-17	1.188968e-13	29.13298
6608	-1.865595	3.514151	-8.667752	6.490934e-16	3.113476e-12	25.58467
284656	2.012712	2.773596	8.655273	7.056437e-16	3.113476e-12	25.50417
4440	-3.471525	-0.179011	-8.620480	8.904162e-16	3.142991e-12	24.85859
57194	1.726910	3.992145	8.546432	1.458409e-15	4.289910e-12	24.78482
100130776	1.428846	2.808023	8.377096	4.470358e-15	1.072779e-11	23.70809
4139	-2.229435	2.442346	-8.358697	5.045362e-15	1.072779e-11	23.55334
55240	1.423331	6.216241	8.346382	5.470569e-15	1.072779e-11	23.44012
10642	-5.268058	-1.255401	-8.117831	2.427554e-14	4.284390e-11	21.61173
79148	2.800647	3.350359	7.982968	5.786983e-14	9.284952e-11	21.21103
355	1.335236	4.408981	7.723453	3.009462e-13	4.426166e-10	19.56352
163732	-1.619732	2.170315	-7.694279	3.615288e-13	4.798675e-10	19.42615
348093	-1.689730	1.716219	-7.676076	4.052791e-13	4.798675e-10	19.28968
202374	2.126031	1.648510	7.669520	4.222851e-13	4.798675e-10	19.29343

Figure 5: Table of the 15 top-ranked genes sorted by p-value for DNAm age acceleration of cancer samples.

No gene was significantly differentially expressed correlated with DNAm age acceleration for the healthy sample data set. For tumors samples, 4503 genes are found differentially expressed, 2159 genes have their expression positively correlated with DNAm age acceleration and 2344 genes have their expression negatively correlated as summarised in table 12. The significant threshold is defined using an adjusted p-value cutoff that is set at 5 percent by default.

Up	2159
Down	2344
Not significant	13146

Table 12: Summary of the number of genes over-expressed (Up), under-expressed(Down) and not significantly differentially expressed for DNAm age acceleration of cancer samples.

It would be interesting to further explore the properties of the main genes statistically significant and up or down-regulated as the gene with the ID 644168. In figure 6, it appears highly up-regulated and the most significant for DNAm age acceleration. This gene is the dorsal root ganglia homeobox in Homo sapiens.

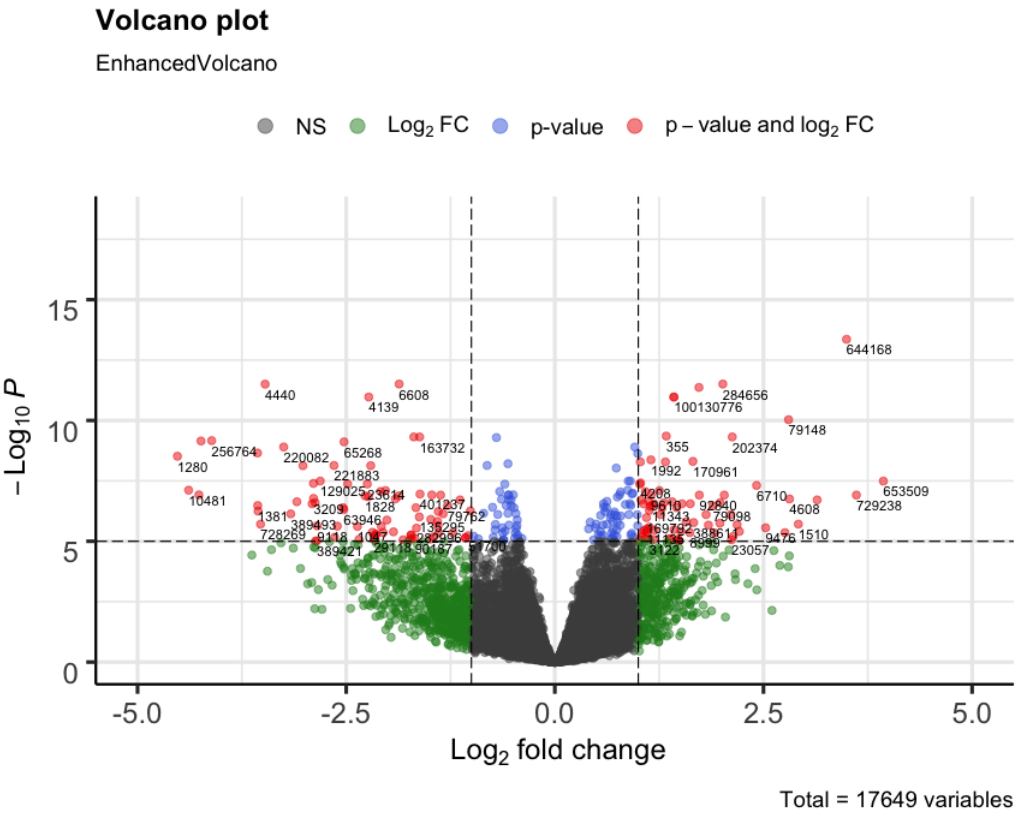

Figure 6: Volcano plot of differentially expressed genes for DNAm age acceleration of cancer samples.

4 Complementary questions.

4.1 Correlation between DNAm age acceleration and the number of somatic mutations for tumor samples.

The number of somatic mutations by samples are obtained from 'Mutation_Packager_Calls', a Mutation Annotation Format (MAF) files from TCGA data. This MAF file store somatic variants of LUAD samples. There is no statistically significant correlation between the DNAm age acceleration and the number of somatic mutations. The result of the spearman correlation is detailed in table 13 and is obtained over 134 samples. This correlation was also not statistically significant in Horvath's study[3].

	rho	p-value
spearman correlation	-0.0710821	0.4144

Table 13: Result of the spearman correlation between DNAm age acceleration and the number of somatic mutations for tumor samples.

4.2 Correlation between DNAm age acceleration and the number of DNA copy number breakpoints.

The results in this section were generated using the 'genomewidesnp6segmentedscnaminusgermlinecnvhg19' file from LUAD cancer TCGA data. This copy number variation (CNV) file contains 115836 observations, one observation for each segment with the start and end of the given segment, the sample barcode, the chromosome of the given segment, his number of probes and the segment mean. The segment_mean refers to the average Log2 ratio of the probes in a given segment. A segment mean close to zero indicates that the given segment has the same number of copies as the germline (used as reference). If the segment is above a certain threshold then it has more copies than the reference, if the segment is below then it has fewer copies than the reference. The value of this threshold is arbitrary. In this section, the chosen threshold is + -0.3.

Before merging the file containing the DNAm age acceleration with the CNV file by patient number, labels were added as follow:

- labelled 0 for neutral, when segment mean is between -0.3 and 0.3.
- labelled 1 for amplification, when segment mean is greater than 0.3.
- labelled -1 for deletion, when segment mean is smaller than -0.3.

4.2.1 Results.

Three spearman correlations were performed; between the DNAm age acceleration and the number of segments with abnormally high number of copies, the number of segments with abnormally low number of copies and the number of segments with abnormally high or low number of copies. The results of the spearman tests are in table 14. The three correlations are statistically significantly negatively correlated with the DNAm age acceleration.

	rho	p-value
Amplification	-0.2812144	1.205e-05
Deletion	-0.1773439	0.005666
Amplification and deletion	-0.2414205	0.0001865

Table 14: Result of the spearman correlation between DNAm age acceleration and the number of segments with abnormally high or/and low number of copies for tumor samples.

The DNAm age acceleration is moderately negatively correlated with the number of segments with abnormally high number of copies and the number of segments with abnormally high or low number of copies. The more the DNAm age acceleration is high, the less is the number of segments with amplifications and the number of segments with deletion or amplifications as shown in figures 7 and 9, respectively. The DNAm age acceleration is weakly negatively correlated with the number of segments with an abnormally low number of copies.

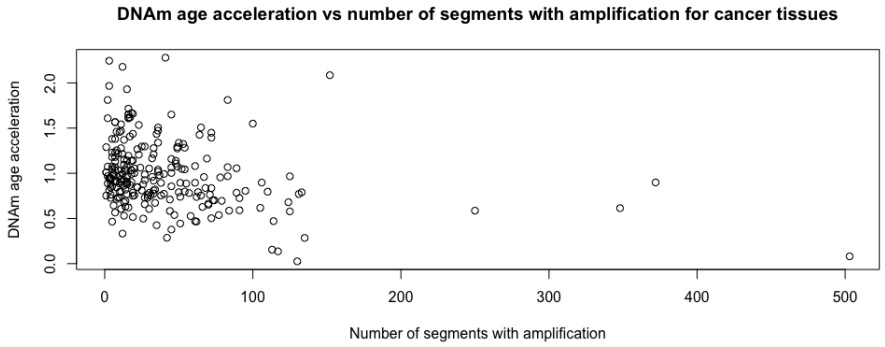

Figure 7: DNAm age acceleration of cancer samples versus the number of segments having abnormally high number of copies.

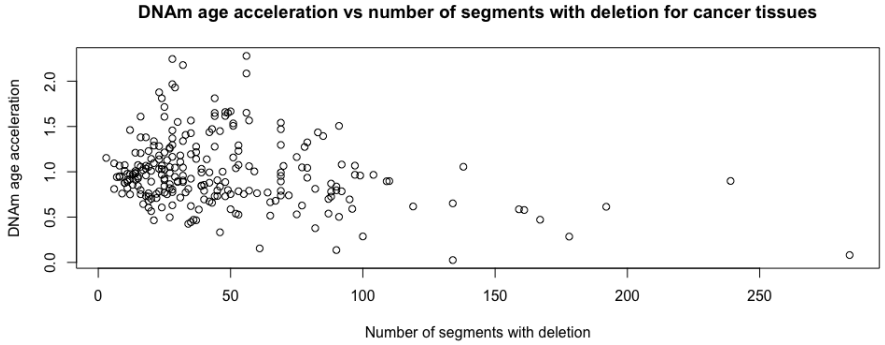

Figure 8: DNAm age acceleration of cancer samples versus the number of segments having abnormally low number of copies.

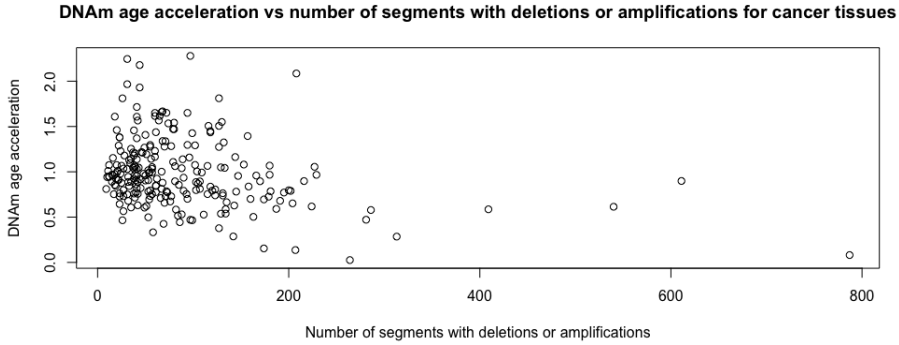

Figure 9: DNAm age acceleration of cancer samples versus the number of segments having abnormally high or low number of copies.

The copy number variation data could be explored in other ways. One possibility would be to take into account the length of the segment based on the number of base pairs. An even better method would be to take into account the number of genes amplified or deleted. This can be

achieved in particular by using the R function Homo.sapiens in order to calculate the number of genes per segment. This last method allows a more detailed and relevant interpretation.

5 Conclusion.

No statistically significant correlation between DNAm age of tumors and their patient-matched normal tissue and between DNAm acceleration for cancer tissues and chronological age were observed. However, a highly positive correlation between DNAm age for healthy tissues and chronological age with a spearman coefficient of 0.76 was found, as expected. Finally, a weak positive correlation between DNAm age for cancer tissues and chronological age (rho = 0.2089718) and significant with a p-value of 0.001101 was observed. Despite a weakly positive but significant correlation between DNAm and chronological age in tumor samples, no statistically significant correlation between the DNAm age acceleration and both quantitative and categorical variables were found. The results obtained by the survival analysis were also not significant. This could be due to the data set size, clinical annotations were available for a maximum of 241 tumours samples and can be much lower for some variable (as the karnofsky performance score for which only 65 samples are annotated).

The correlation between the C2:CP gene expressions and DNAm age acceleration provided 8 C2:CP gene sets statistically significant and specific[10] to tumour samples, by order of statistical significance: SA MMP CYTOKINE CONNECTION, SIG PIP3 SIGNALING IN B LYMPHOCYTES, NABA ECM AFFILIATED, NABA MATRISOME ASSOCIATED, NABA PROTEOGLYCANS, SIG BCR SIGNALING PATHWAY, SA FAS SIGNALING and NABA MATRISOME. The two most significant gene sets are described in the scientific literature as having a concrete role in non-small cell lung cancer[13][14][15] [16].

No gene was significantly differentially expressed correlated with DNAm age acceleration for the healthy sample data set. However, for tumors, 4503 genes are found differentially expressed, 2159 genes have their expression positively correlated with DNAm age acceleration and 2344 genes have their expression negatively correlated. Literature scientific research about these up and down-regulated genes is desirable for future work.

Despite a study published in 2014 suggests that lung adenocarcinoma has a high rates of somatic mutation[6]; no statistically significant correlation between the DNAm age acceleration and the number of somatic mutations in tumor tissues was found. This is in accordance with the study carried out by Horvath [3]. Finally, the DNAm age acceleration is moderately negatively correlated with the number of segments with abnormally high number of copies(rho = -0.2812144). The DNAm age acceleration is weakly negatively correlated with the number of segments with an abnormally low number of copies (rho -0.1773439). Future studies could investigate the association between the number of genes having an abnormal number of copies and the DNAm age acceleration.

In addition, the democratisation and advancement of modern high throughput sequencing techniques, will hopefully increase the data available to study DNA methylation.

[10] 8 out of 9 were specific to tumor, the NABA SECRETED FACTORS gene set expression has a positive correlation with DNAm age acceleration for both tumor and healthy samples.

References

[1] Bilian Jin, Yajun Li, and Keith D Robertson. Dna methylation: superior or subordinate in the epigenetic hierarchy? *Genes & cancer*, 2(6):607–617, 2011.

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

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

[4] Anthony J Alberg and Jonathan M Samet. Epidemiology of lung cancer. *Chest*, 123(1):21S–49S, 2003.

[5] Zhao Chen, Christine M Fillmore, Peter S Hammerman, Carla F Kim, and Kwok-Kin Wong. Erratum: Non-small-cell lung cancers: a heterogeneous set of diseases. *Nature Reviews Cancer*, 15(4):247–247, 2015.

[6] Cancer Genome Atlas Research Network et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, 2014.

[7] Jonathan M Samet, Erika Avila-Tang, Paolo Boffetta, Lindsay M Hannan, Susan Olivo-Marston, Michael J Thun, and Charles M Rudin. Lung cancer in never smokers: clinical epidemiology and environmental risk factors. *Clinical Cancer Research*, 15(18):5626–5645, 2009.

[8] D. J. Best and D. E. 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.

[9] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[10] Charity W Law, Monther Alhamdoosh, Shian Su, Xueyi Dong, Luyi Tian, Gordon K Smyth, and Matthew E Ritchie. Rna-seq analysis is easy as 1-2-3 with limma, glimma and edger. *F1000Research*, 5, 2016.

[11] Charity W Law, Yunshun Chen, Wei Shi, and Gordon K Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):R29, 2014.

[12] 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.

[13] Pragya Misra and Shailza Singh. Role of cytokines in combinatorial immunotherapeutics of non-small cell lung cancer through systems perspective. *Cancer medicine*, 8(5):1976–1995, 2019.

[14] Lindsey Enewold, Leah E Mechanic, Elise D Bowman, Yun-Ling Zheng, Zhipeng Yu, Glenwood Trivers, Anthony J Alberg, and Curtis C Harris. Serum concentrations of cytokines and lung cancer survival in african americans and caucasians. *Cancer Epidemiology and Prevention Biomarkers*, 18(1):215–222, 2009.

[15] Si-si Wang, Wei Liu, Dalam Ly, Hao Xu, Limei Qu, and Li Zhang. Tumor-infiltrating b cells: their role and application in anti-tumor immunity in lung cancer. *Cellular & Molecular Immunology*, 16(1):6–18, 2019.

[16] Kei Suzuki, Stefan S Kachala, Kyuichi Kadota, Ronglai Shen, Qianxing Mo, David G Beer, Valerie W Rusch, William D Travis, and Prasad S Adusumilli. Prognostic immune markers in non–small cell lung cancer. *Clinical Cancer Research*, 17(16):5247–5256, 2011.

Contents