BIOL-F-423: PROJECT 2016

Professor: Vincent DETOURS

Student : Joel RODRIGUES VITORIA

25/05/2016

CONTENTS

1. INTRODUCTION

One of the main research branches in the relatively young field of bioinformatics is the handling, analysis and interpretation of genomic data. Said data, as become widely accessible since the rise of various Next Generation Sequencing (**NGS**) technologies at the beginning of the 21th century (Barba et al., 2014) . This project aims at familiarizing the students with the type of data produced by different sequencing platforms and technics, more specifically in form of publicly available data (raw or pre-processed).

We will be using data produced by The Cancer Genome Atlas (**TCGA**), a public founded project, launched by the National Institute of Health (**NIH**) back in 2005. Today, this enormous collaborative effort has produced matched tissue data from 11.000 patients and characterizes 33 cancer types[1] (Tomczak et al., 2015).

The main objective of this project is the corroboration of some observations done by Steve Horvath, a bioinformatician that developed a multi-tissue predictor of age based on DNA methylation (Horvath, 2013). Among many interesting findings, Horvath noted that all cancers presented a positive DNAm age acceleration, defined as the mean difference between DNAm age and chronological age. This statement has since then been corrected by the author in an erratum (Horvath, 2015), and it seems that out of 20 cancer types, only 6 exhibit positive age acceleration while most others present negative age acceleration. Nevertheless, it seems that certain somatic mutations seem correlated with DNAm age acceleration, *e.g.* TP53 mutations.

In light of these findings, our objective will be to explore possible DNAm age correlations based on a random patient sample, for which the DNAm age was calculated based on TCGA data.

2. METHODOLOGY

2.1. DATA DESCRIPTION

All original files were either supplied by the professor of this course (Vincent Detours) or obtained through the firehose access interface for the TCGA data[2]. The author of this document was given a dataset based on COloREctal Adenocarcinoma (COADREAD) tissue samples.

- ***COADREAD-8.Rda:*** a data object used for the statistical programming language R[3]. It contains the calculated DNAm age for our randomly chosen patient samples.
- ***COADREAD.clin.merged.picked.txt:*** a tab-separated text file containing the retained clinical variables used by the TCGA team.

1 http://cancergenome.nih.gov/abouttcga/overview
2 http://firebrowse.org/?cohort=COADREAD&download_dialog=true
3 https://www.r-project.org/

- ***COADREAD.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes__data.data.txt:*** *tab-separated* text file containing the expression data for genes obtained from RNA-Seq analysis. Raw counts are given in this file.
- ***COADREAD.rnaseqv2__illuminahiseq_rnaseqv2__unc_edu__Level_3__RSEM_genes_normalized__data.data.txt:*** *tab-separated* text file containing the expression data for genes obtained from RNA-Seq analysis. Count values were normalized.

All other files were generated during the analysis pipeline and will be described as needed. All R language scripts, objects and results were attached to this report as a zipped directory (Project_2016).

2.2. USED SOFTWARE

As stated above, the open source language R *version 3.3.0* was used. Additionally an graphical interface for R was used, Rstudio *version 0.99.893,* which other that providing a visual working environment for script writing and plotting, does not modify the base R distribution in any way.

While most statistical analyses use R implementations (*library packages*), we used the java application for the Gene Set Enrichment Analysis (**GSEA**), which is distributed and maintained by the Broad Institute[4].

All R scripts provided should work as is, as the script will try to download any needed library package, for as long as an Internet connection is provided. The three scripts, one per project question, used the directory tree of the main project folder, thus functionality cannot be guaranteed if folder names are modified.

Output of the *sessioninfo()* function are given in the annexes and show the working environment used during the execution of the scripts.

2.3. STRUCTURE OF REPORT

Since the project is build around 3 major questions, the rest of this document will be structured around these. Thus, for each question we will present the specific data treatment methodology and the results/conclusions.

4 http://software.broadinstitute.org/gsea/downloads.jsp

3. Q1: HOW DOES CHRONOLOGICAL AGE AND DNAM AGE CORRELATE?

3.1. METHODOLOGY

As requested, scatter plots and Spearman correlations were used in this part. For this, the original data was formatted into a main *data.frame* object (which basically is a data table). *See question_one.R script file.*

Six columns where produced:

- **sample and DNAm_age** : contains respectively the TCGA sample barcode and calculated DNAm age. Both were extracted from *dm.age.subset*, the named vector that is loaded into R when loading the *COADREAD-8.Rda* file.
- **tissu:** dichotomous tissue attribution, taking two possible values (*normal* or *tumor*). For this we used the official TCGA barcode specifications[5] and extracted the sample type field (numeric value [01-29]) with a regular expression. The fields where mapped with the following rule: if less than 9 return "tumor" else return "normal. No finer distinction between primary/metastasis and blood/normal tissues was applied, but after checking the regular expression matches, only samples with either code 01 (primary tumor) or code 11 (solid normal tissue) were present.
- **patientID:** contains the three first fields from the barcode, which is the unique patient identifier.
- **years_to_birth:** contains chronological age. These values where extracted from the clinical variables text file by mapping them to the patientIDs.
- **paired:** Boolean true/false column. It was created by using two conditions for a TRUE value: 1) a given patientID is present exactly two times; 2) These two samples represent each a matched tumor/normal sample. 168 samples passed this filter, giving us 84 matched tumor/normal samples. Three patientIDs did not pass the first filter (TCGA.A6.2677, TCGA.A6.3809, TCGA.A6.3810). It was decided to exclude them instead of, *e.g.* calculating an average DNAm age value for the two tumor samples present, given their small number.

This main data frame was subset when needed and plots where created using the *ggplot2* package.

5 https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode

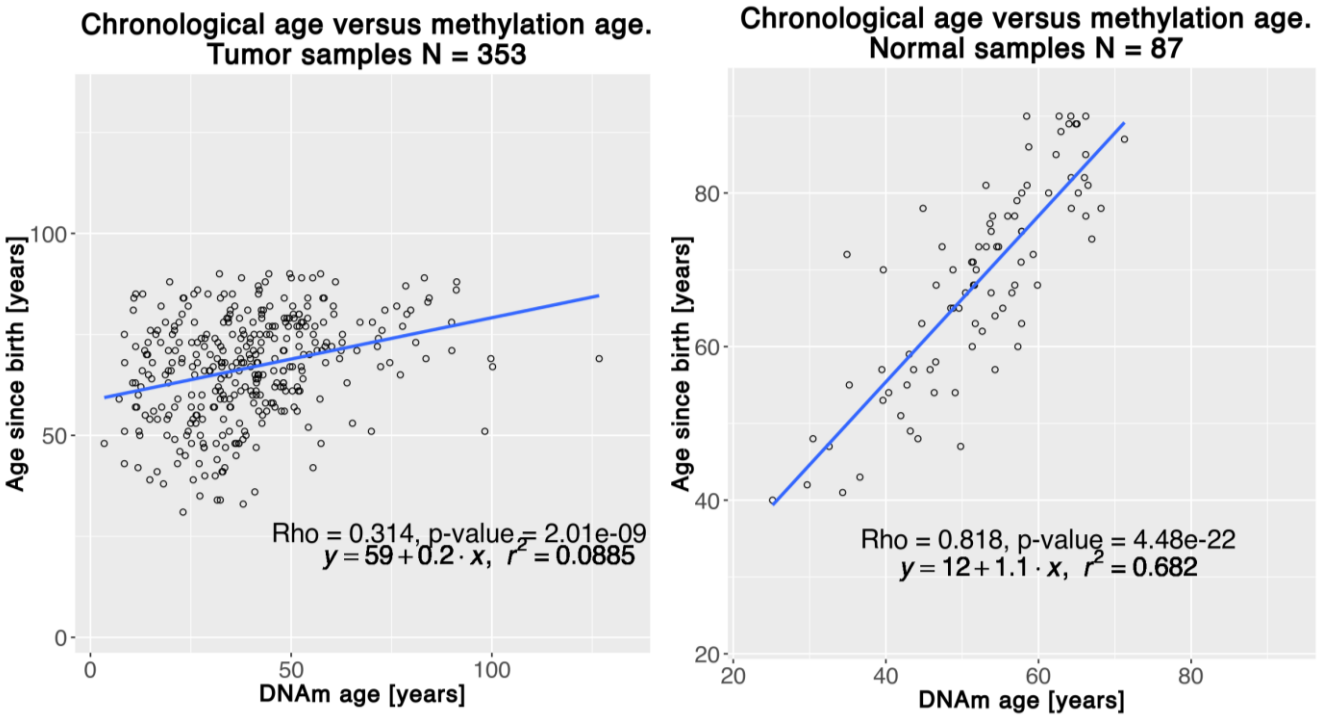

Figure 1: scatter plots showing chronological age versus calculated DNA methylation age. Blue line shows the linear regression as given by the formula in the plot. Spearman rho and p-value are given. **LEFT:** data for tumor tissue samples. **RIGHT:** data for normal tissue samples

We plotted the chronological age vs. the calculated DNAm age by subsetting the data by tissue type (Figure 1). The Pearson correlation coefficient suggest a better correlation for the normal tissue samples (Rho = 0.818, p-value = 4.48e-22) than for the tumor tissue samples (Rho = 0.314, p-value = 2.01e-9). Subsequently, the r-squared values are better in the normal tissue linear regression (r^2 = 0.682) than the one for tumor tissue samples (r^2 = 0.0885). This suggests that DNAm age is a better estimator for normal tissue chronological age than for tumor samples.

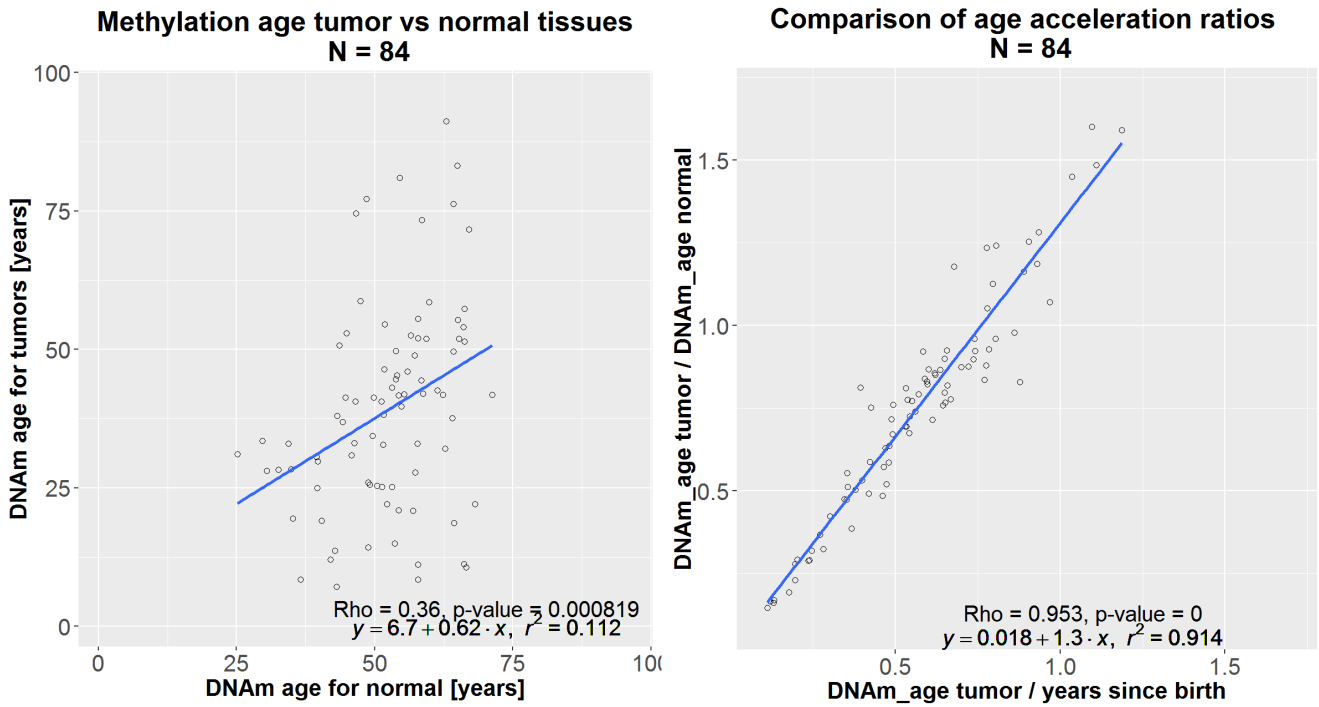

Figure 2: **LEFT:** scatter plot showing the correlation between patient-matched tumor/normal samples. **Right:** scatter plot showing the correlation between two definitions of the DNAm age acceleration, i.e. either tumor DNAm age divided by the matched normal tissue DNAm age, or the tumor DNAm age divided by the chronological age of the patient. Blue line shows the linear regression as given by the formula in the plot. Spearman rho and p-value are given.

When comparing the calculated DNAm age of patient matched tissue samples (Figure 2), we observe a slight positive correlation (Rho = 0.36, p-value = $8*10^{-4}$), suggesting that both are marginally correlated, and looking at the linear correlation formula, the slope (0.62) indicates that the DNAm ages for the tumor samples are systematically smaller than the chronological age of the patient.

On the other hand, both DNAm age acceleration definitions are strongly correlated (Rho = 0.953, p-value < 0.001), suggesting that both definitions, tumor DNAm age divided by the DNAm age of normal tissue or divided by the chronological age, bear approximatively the same information.

3.3. CONCLUSION

Our results show that the DNAm age calculated for the normal tissues have a slight positive correlation with the chronological age of the patients, which corroborates its value as an age predictor. But, we also observe that this conclusion can't be drawn for the tumor samples, which implies that the methylation process is modified in tumor cells.

Indeed, most tumor samples show a reduced DNAm age acceleration (i.e. a value smaller than 1), leading us to believe that our tumor cells appear “younger” than the matched normal tissue. Another important result is the fact that both DNAm age acceleration definitions are equivalent in terms of information they bear, thus we will prefer the definition based on chronological age for the rest of the analysis. This is mostly based on the larger sample size at our disposal when using this definition, since we will be able to analyze all tumor samples (not just the tumor/normal patient-matched samples).

4. Q2: ARE THERE CLINICAL VARIABLES CORRELATED WITH DNAM AGE IN CANCERS?

4.1. METHODOLOGY

To study the possible clinical variables correlated with DNAm age, the first part consisted in generating a main data frame for easier data handling. *See question_two.R script file.*

The generated data frame (*dfQuestionTwo*) is based on the one from question one, to which where appended a number of columns representing the rest of the information available in the *COADREAD.clin.merged.picked.txt,* after cross-matching the patient IDs with those present in our sample. We will briefly describe the different clinical variables for later reference (all but the ones already described in question one).

- **vital_status:** dichotomous vector indication death (value 1) or alive status (value 0). These are right censored survival statuses, meaning that the alive status is only valid up to the last follow up date (see below).
- **days_to_death:** number of days starting from day of diagnosis up to the death of the patient.
- **days_to_last_followup:** number of days starting from day of diagnosis up to the last medical followup.
- **tumor_tissue_site:** location of the tumor samples indicated by either *colon* or *rectum*.
- **pathologic_stage:** character vector containing 12 different pathological stages, ranging from stage 0 up to stage IVa. These increasing stages describe the severity of the cancer spread in the body.
- **pathology_T_stage:** character vector containing 7 different pathological stages for primary tumor, based on the American Joint Committee on Cancer (**AJCC**, 7e edition).
- **pathology_N_stage:** character vector containing 9 different pathological stages for tumor spread in local lymph nodes based on the American Joint Committee on Cancer (**AJCC**, 7e edition).
- **pathology_M_stage:** character vector containing 4 different pathological stages indicating if there is metastasization in other parts of the body, based on the American Joint Committee on Cancer (**AJCC**, 7e edition).
- **gender:** character vector indicating the gender of the patient (*male* or *female*).
- **date_of_initial_pathologic_diagnosis:** vector indicating the year of first diagnosis of cancer.
- **days_to_last_known_alive:** days from initial diagnosis up to last known date of known alive status.
- **radiation_therapy:** character vector indicating if a radiation therapy was used (*yes* or *no*).
- **histological_type:** character vector containing 4 different histological types of tumors (*colon adenocarcinoma, colon mucinous adenocarcinoma, rectal adenocarcinoma* and *rectal mucinous adenocarcinoma*).

- **tumor_stage:** vector supposed to indicate tumor pathological spread. *This vector was empty in our case (NA values), thus it was not used in our later analyses.*
- **residual_tumor:** text term to describe the status of tissue margins following surgical resection for a locoregional procedure.
- **number_of_lymph_nodes:** integer indicating number of lymph nodes involved with the disease
- **ethnicity:** character vector indicating the ethnic group of the patient. In our case this vector either indicates *not hispanic or latino* or an *NA* value.
- **age_acc:** DNAm age acceleration calculated as DNAm age of tumor divided by chronological age of patient.

For further information on all the possible clinical variables for the TCGA project, visit the official documentation[6].

For nominal/categorical data, the following statistical procedure was used: after generating boxplots, a Kruskal-Wallis non-parametric test was applied. If the p-value was significant (alpha = 0.05), we applied a pair-wise multiple comparison as to determine which pairs of variables were significantly different. For this, we used the *PMCMR* package, which contains the *posthoc.kruskal.nemenyi.test* function, a *posthoc* method based on the works of Nemenyi[7]. It uses a *family-wise error* method for the p-value correction of multiple tests. The significant pairs where manually annotated on the boxplots.

For the numerical (discrete) data, namely the date of first diagnosis and the number of lymph nodes, we used Spearman correlation to assess a significant correlation.

For the analysis of survival data we started by first concatenating the *days_to_death* and *days_to_last_followup* columns. Since both vectors are mutually exclusive, meaning that a *NA* value in one vector implies a numerical value in the other one, we converted the *NA* values to zero values and did a vector sum to create a *surv_data* column. This column, in addition to *vital_status,* was used to apply a Cox proportional hazards regression model on the data.

6 https://tcga-data.nci.nih.gov/docs/dictionary/
7 https://cran.r-project.org/web/packages/PMCMR/vignettes/PMCMR.pdf

4.2.1. DNAM AGE

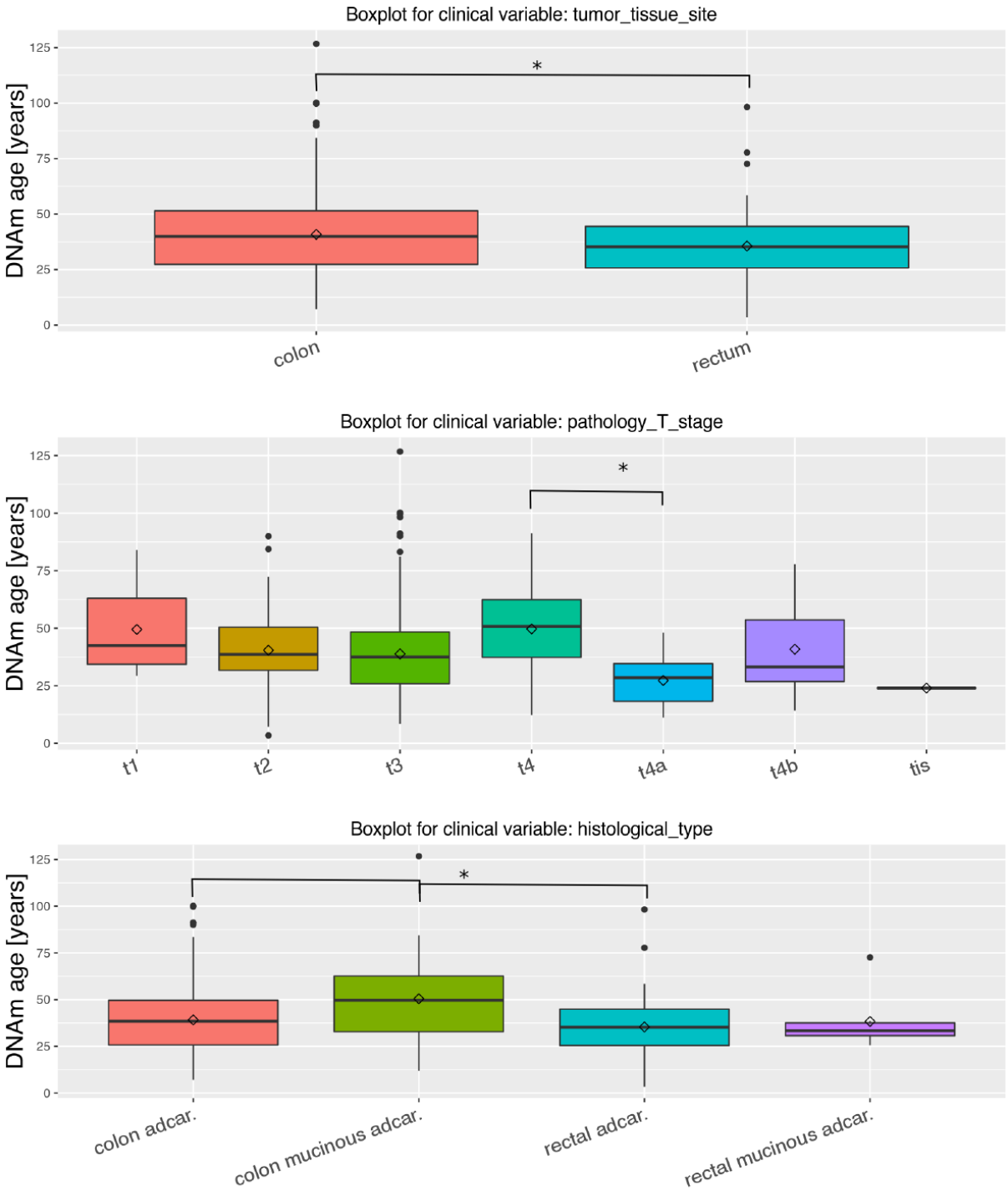

Figure 3: boxplots for DNAm age versus the clinical variables. Only plots that showed a significant p-value (alpha = 0.05) when using a Kruskal-Wallis test, are represented. Bars with a star, show significant differences after applying a posthoc multiple comparisons test (after Nemenyi). For detailed description of clinical variables, see methodology. Square-shaped point shows the calculated mean value.

Table 1: Summary of the Kruskal-Wallis test results for nominal/categorical clinical variable. Test variable = DNAm age.

Clinical factor	Sample size	chi-sqr	df	p-value
tumor_tissue_site	350	4.6988	1	0.0302
pathologic_stage	343	14.6841	11	0.1974
pathology_T_stage	350	14.8596	6	0.0214
pathology_N_stage	350	14.1737	8	0.0773
pathology_M_stage	347	3.2700	3	0.3518
gender	352	1.2645	1	0.2608
radiation_therapy	294	0.7978	1	0.3717
histological_type	347	12.0669	3	0.0072
residual_tumor	297	3.7406	3	0.2909
ethnicity	190	2.4464	1	0.1178

After applying the Kruskal-Wallis tests (Table 1), only three clinical variables returned a significant p-value: tumor tissue type, pathology T stage and histological type. After applying pairwise multiple tests, we observe that colon tumors have higher mean DNAm age than rectum tumors (40.85 +-19.55 versus 35.61 +-15.60 years; Figure 3). Among the pathological tumor stages, only stage t4 and t4a show a significant difference (49.69 +-22.13 versus 27.30 +-11.97 years). Finally, colon mucinous adenocarcinomas show higher mean DNAm age (50.47 +-23.03 years) than colon adenocarcinomas (39.15 +-18.38 years) and rectal adenocarcinomas (35.37 +-15.67 years).

There was no significant Spearman correlation for the quantitative discrete clinical variable *date_of_initial_pathological_diagnosis*. *Number_of_lymph_node* has a significant p-value (0.00492) but a small r^2 values (0.0182; Figure 4).

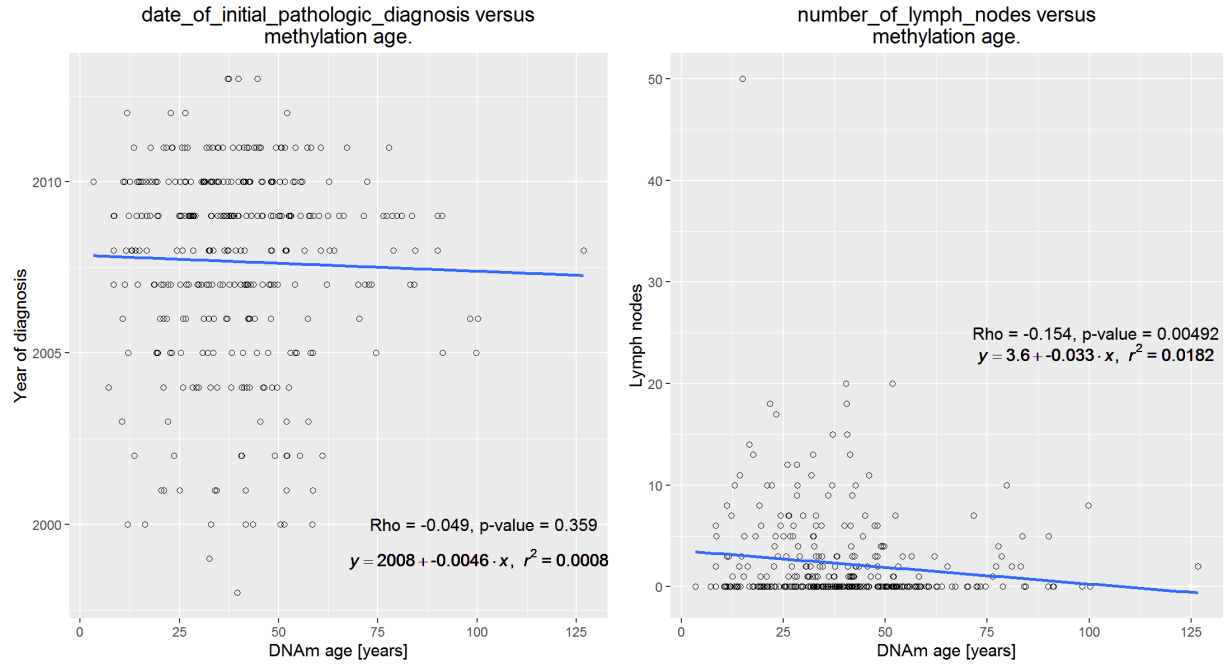

Figure 4: scatter plots showing correlation for quantitative (discrete) clinical variables versus DNAm age. Blue line shows the linear regression as given by the formula in the plot. Spearman rho and p-value are given.

4.2.2. DNAM AGE ACCELERATION

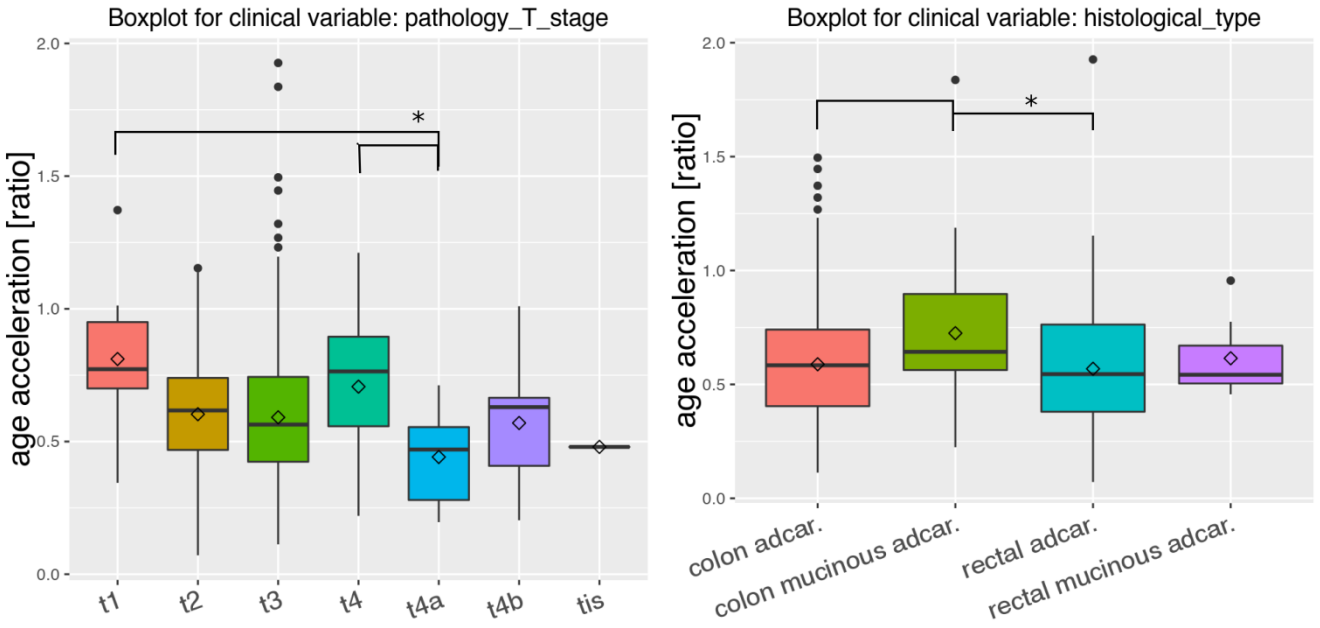

Figure 5: boxplots for DNAm age acceleration versus the clinical variables. Only plots that showed a significant p-value (alpha = 0.05) when using a Kruskal-Wallis test, are represented. Bars with a star, show significant differences after applying a *posthoc* multiple comparisons test (after Nemenyi). For detailed description of clinical variables, see methodology. Square-shaped point shows the calculated mean.

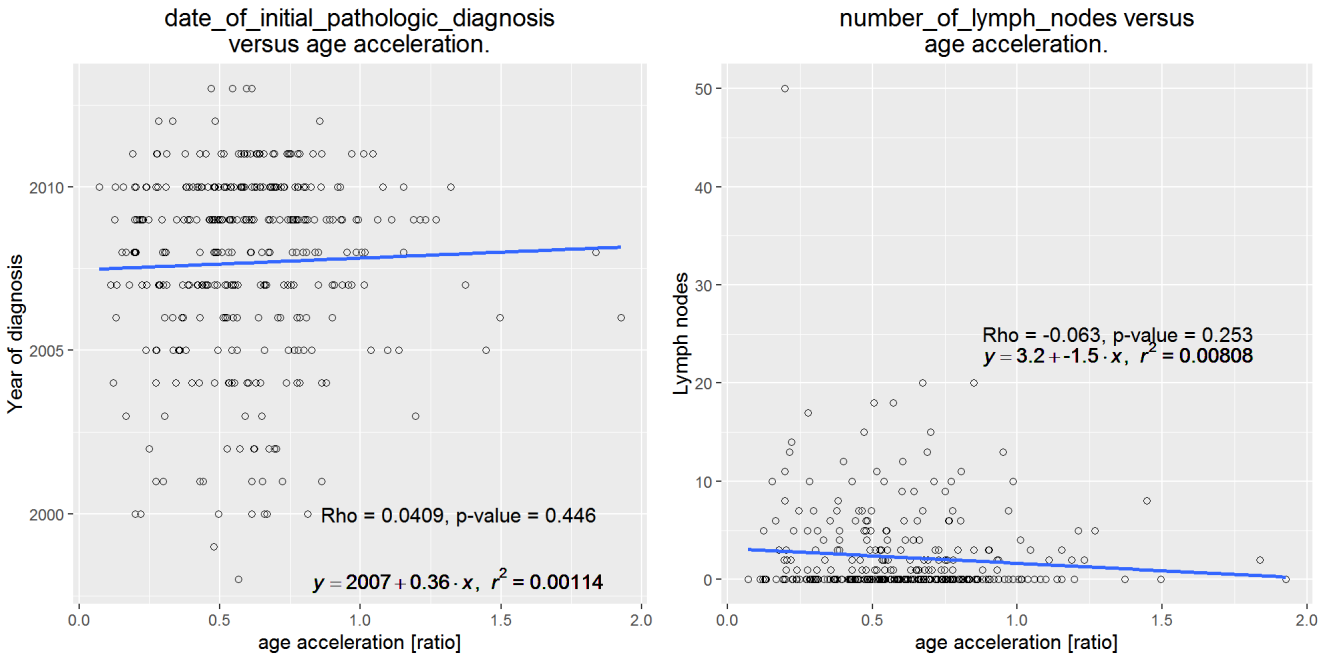

Figure 6: scatter plots showing correlation for quantitative clinical variables versus DNAm age acceleration. Blue line shows the linear regression as given by the formula in the plot. Spearman rho and p-value are given.

After applying the Kruskal-Wallis tests (Table 2), only two clinical variables returned a significant p-value: pathology T stage and histological type. Among the pathological tumor stages, stage t4a shows a

significant lower mean age acceleration (0.44) compared to t4 (0.71) and t1 (0.81; Figure 5). Finally, colon mucinous adenocarcinomas show significantly higher mean DNAm age acceleration (0.72) than colon adenocarcinomas (0.59) and rectal adenocarcinomas (0.56).

There were no significant Spearman correlations for the quantitative clinical variables (*date_of_initial_pathological_diagnosis* and *number_of_lymph_node*; Figure 6).

Table 2: Summary of the Kruskal-Wallis test results for nominal/categorical clinical variable. Test variable = DNAm age acceleration.

Clinical factor	Sample size	chi-sqr	df	p-value
tumor_tissue_site	350	0.9226	1	0.3368
pathologic_stage	343	8.4461	11	0.6729
pathology_T_stage	350	14.7121	6	0.0226
pathology_N_stage	350	9.3698	8	0.3121
pathology_M_stage	347	2.1803	3	0.5358
gender	352	1.3215	1	0.2503
radiation_therapy	294	0.4122	1	0.5208
histological_type	347	7.9037	3	0.0480
residual_tumor	297	2.9460	3	0.4000
ethnicity	190	0.0893	1	0.7650

4.2.3. COX REGRESSION ANALYSIS

We tested two Cox regression models:

- survival ~ age_acc + DNAm_age + interaction(age_acc:DNAm_age)
- survival ~ age_acc + DNAm_age

Table 3: R output when applying the Cox proportional hazard regression model (*coxph()* function) based on our two regression models.

```
Call:
coxph(formula = Surv(surv_data, vital_status) ~ age_acc + DNAm_age +
    age_acc:DNAm_age, data = cancerdf)

  n= 350, number of events= 80
   (2 observations deleted due to missingness)

                       coef     exp(coef)        se(coef)     z    Pr(>|z|)
age_acc                -1.87812  0.15288  1.35977    -1.381   0.1672
DNAm_age               0.03813   1.03887  0.01874     2.035   0.0419 *
age_acc:DNAm_age       -0.01069   0.98936  0.01628    -0.657   0.5113
---
Signif. Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Call:
coxph(formula = Surv(surv_data, vital_status) ~ age_acc + DNAm_age,
    data = cancerdf)

  n= 350, number of events= 80
```

```
  (2 observations deleted due to missingness)

                coef     exp(coef)       se(coef)     z     Pr(>|z|)
age_acc         -2.38124   0.09244      1.14983 -2.071   0.0384 *
DNAm_age        0.03125   1.03174       0.01566  1.996   0.0459 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The interaction between age acceleration and DNAm age was not significant (p-value = 0.5113, Table 3). Thus, we proceeded to a simpler model only taking into account the age acceleration and the DNAm age. Both variables produced a significant p-value (0.0384 and 0.459). The age acceleration shows an exp(coef) of 0.09244, which means that increasing the age acceleration by one unit reduces the risk of death by about 91%. For the DNAm age, increasing by one unit increases the death risk by about 3% (exp(coef) = 1.03174).

4.3. CONCLUSIONS

The data analysis suggests that there are some correlations for DNAm age and age acceleration when tested against the clinical variables. Most notably, the primary tumor stage 4a showed consistent lower mean DNAm age and low age acceleration, suggesting that this stage is populated predominately by cancer cells that appear “younger” in terms of methylation ages. Thus, one could propose a hypothesis by which a late stage cancer shows signs of a modification in methylation of DNA, giving them properties of less older cells. Nevertheless, one should be careful, since our data only points to such a conclusion and further study of the underlying mechanisms that govern the DNA methylation is needed as to stipulate a more robust hypothesis in terms of the biological implications.

The colon mucinous adenocarcinomas seem consistently older and have higher age acceleration ratios. It would be interesting to verify if these histological types seem less aggressive than their other counterparts, which could help us interpret the effects of the DNAm age predictor in terms of symptoms.

Also, for the DNAm age there was a slight negative correlation with the number of lymph nodes involved in the disease, suggesting that a lower DNAm age phenotype is related to more aggressive tumors. But, not only was this correlation weak, the same results could not be replicated for the DNAm age acceleration.

Lastly, the Cox regression clearly shows that a low age acceleration ratio and a higher DNAm age increase the risk death risk. Here again, it appears that the “younger” phenotype of cancer cells is correlated with potentially more aggressive cancers.

5. Q3: WHICH GENES/PATHWAYS HAVE EXPRESSION ASSOCIATED WITH DNAM AGE IN CANCER?

5.1. METHODOLOGY

In this section, as a way to study the genes and pathways associated with DNAm age, we chose to use 2 different analyzing algorithms:

1. Gene Set Enrichment Analysis (**GSEA**): this method determines if a given, predefined gene set is significantly overrepresented in a given sample (Subramanian et al., 2005). It calculates an enrichment score, for which p-values and false discovery rates (**FDR**) are calculated and based on which, significant pathway gene set are isolated.
2. Significance Analysis of Microarrays (**SAM**): a gene selection procedure that uses a moderated t-statistic, on which a permutation-based estimate of the null distribution is used (Tusher et al., 2001). Based on this, a FDR is calculated for the significant expression of genes.

See question_three.R script for details on data treatment pipeline.

Because both methods needed different pre-processing of the raw data, we briefly describe each approach.

5.1.1. GSEA METHODOLOGY

Instead of using an R implementation of the GSEA method, we opted for the java application distributed by the Broad Institute. Due to compatibility issues for the data formats expected by the application, all input files were generated with R after studying the official format specifications.

1. Based on the TCGA normalized RNAseq data, we created a data frame. The data was log2 transformed using the formula *yi= log2(xi + 8).* The constant was added as to avoid infinite values when the normalized count was 0. Also, before log transformation, rows containing only 0 values (*i.e.* rowsum = 0 for a given gene) were removed from the data frame.
2. We subsetted for tumor samples and extracted a continuous co-variate vector for the DNAm age and age acceleration, by matching the patientIDs of our samples. These where saved as a continuous phenotype object for GSEA (*./data_output/question_three/rnaseq_data_log_cleaned.cls*).
3. An Entrez/LocuLink ID column was created by string splitting the original row names which were structured as follows "ID|GENE SYMBOL". This final data frame was then saved as tab-separated text file (*./data_output/question_three/rnaseq_data_log_cleaned.txt*).
4. We generated our own .chip file, which is a three column text file that maps each Entrez/LocuLink ID to the corresponding GENE SYMBOL and gene description (*./data_output/question_three/GSEA_COREAD.chip*).

5. We downloaded the C2:CP(canonical pathways) collection from the MsigDB site[8]. Note that we used the ENTREZ ID version of the file (*c2.cp.v5.1.entrez.gmt*).

All this files were loaded into GSEA v2.2.2 and the following non-standard parameters where used:

- **Expression data set** = rnaseq_data_log_cleaned
- **Gene sets database** = *c2.cp.v5.1.entrez.gmt*
- **phenotype labels** = *rnaseq_data_log_cleaned.cls#DNAm_age* **or** *rnaseq_data_log_cleaned.cls#DNAm_acc*
- **Collapse dataset to gene symbols** = FALSE
- **Chip platform** = *GSEA_COREAD.chip*
- **Metric for ranking genes** = Pearson

The program generated *html* reports that are located in the "*Project_2016/GSEA_output/*" directory.

5.1.2. SAM METHODOLOGY

We used the *samr* package in R for this part. More specifically the *SAMseq* function was used, which allows the application of the SAM method on RNAseq data. But, the function expected raw counts as input, thus we had to download the raw RNAseq counts and preprocess the data:

1. The raw text file was loaded as a *data frame* and cleaned. All steps are identical to the GSEA pipeline, except the log transformation of the data.
2. The original data file contains some hypothetical locus IDs (e.g. IDs *100130426, 100133144* …), which for some unknown reason have non-integer raw counts. This became a problem later one, since *SAMseq* expects integer count values. Thus, the count values were rounded has to eliminate this compatibility problem.
3. The *SAMseq* function was run on the raw counts by specifying one of both co-variate vectors (DNAm age or DNAm acceleration) and specifying it as a quantitative co-variate.
4. For both co-variates, tables for up and down regulated genes were generated after filtering them for a FDR < 0.1. These files were saved as csv files in the *Project_2016/data_output/question_three* directory

[8] http://software.broadinstitute.org/gsea/msigdb

5.2. RESULTS

One visual tool to summarize the GSEA results are *normalized enrichment scores(**NES**) versus significance* plots (Figure 7). When using DNAm age as co-variate, we observe numerous significant (FDR < 0.250) gene sets for high NES values, suggesting that some gene sets are upregulated for samples with a positive correlation profile for DNAm age. For DNAm age acceleration, only a small number of gene set seem correlated, and this time they are for negative NES values. This is corroborated by the GSEA summary report (Figure 8), showing 130 significant (based on FDR) gene sets for the positive DNAm age correlation phenotype and 1 significant gene set for negative DNAm age acceleration correlation phenotype.

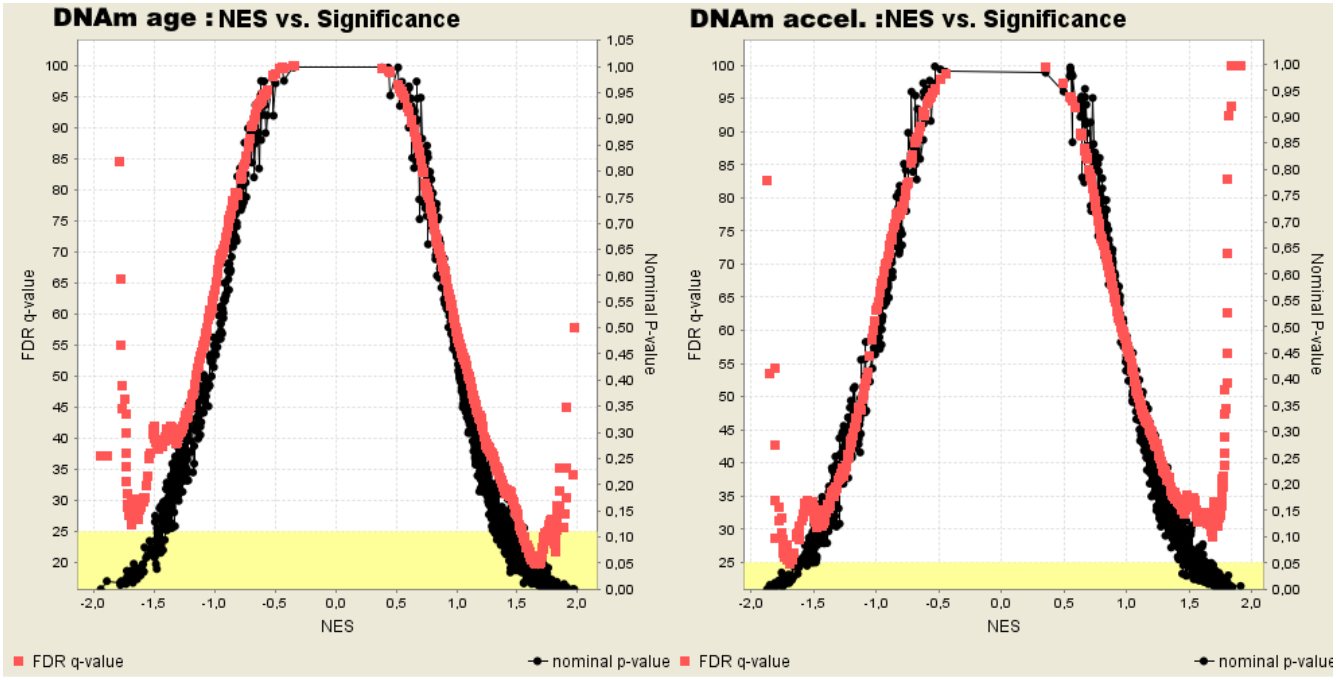

Figure 7: normalized enrichment scores versus significance statistics, as produced by GSEA. The yellow area shows the gene set points that satisfy a FDR < 0.250.

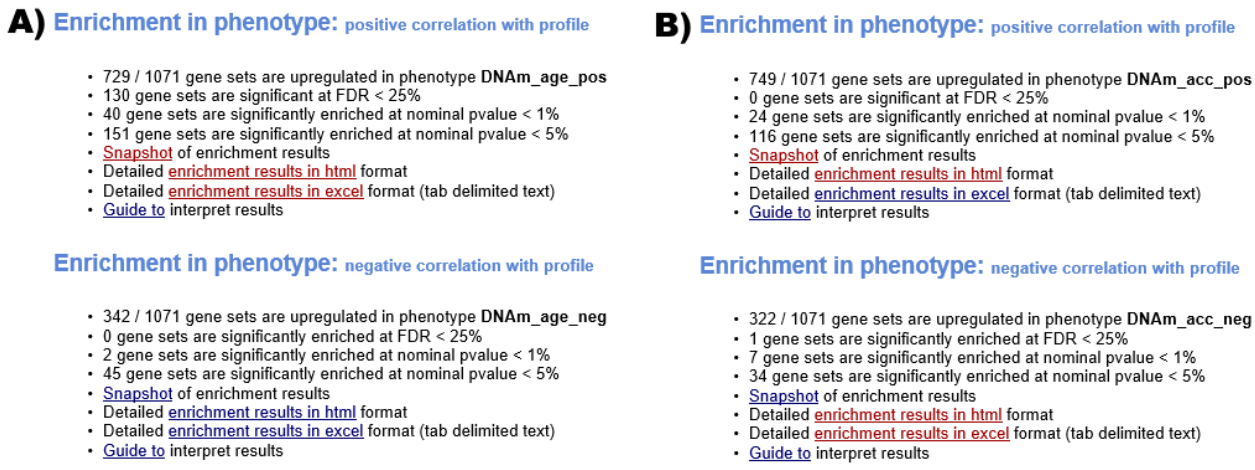

Figure 8: screenshot of the generated html reports from GSEA. **A:** summary for DNAm age as continuous co-variate. **B:** summary for DNAm age acceleration as continuous co-variate.

Since analyzing the whole list of enriched gene set would take too much time and requires an extensive expertise in colon-rectal cancer biology, we will restrict our commentary to a top ten list of significant gene set (based on FDR values).

Table 4: top ten results from GSEA for DNAm age as co-variate. Shown are gene sets, their size, enrichment score (ES), normalized enrichment score (NES), nominal p-value, FDR q-value and max rank obtained during GSEA analysis.

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	RANK AT MAX
BIOCARTA_EPO_PATHWAY	19	0.544	1.683	0.006	0.197	2717
KEGG_TYPE_I_DIABETES_MELLITUS	41	0.641	1.661	0.060	0.198	4165
PID_IL8_CXCR2_PATHWAY	34	0.579	1.646	0.017	0.198	3554
PID_LYSOPHOSPHOLIPID_PATHWAY	66	0.495	1.651	0.029	0.199	4078
KEGG_ALDOSTERONE_REGULATED_SODIUM_REABSORPTION	42	0.518	1.684	0.008	0.199	5241
SA_MMP_CYTOKINE_CONNECTION	15	0.730	1.654	0.021	0.199	5019
REACTOME_CIRCADIAN_CLOCK	52	0.487	1.655	0.010	0.200	5668
PID_ARF6_PATHWAY	35	0.539	1.658	0.022	0.200	2382
BIOCARTA_TH1TH2_PATHWAY	19	0.700	1.661	0.046	0.200	4165

The significant gene sets for DNAm age don't seem to fit a clear biological effect (Table 4). The most significant gene set is related to the erythropoietin functions (*BIOCARTA_EPO_PATHWAY*), while others are related to interleukin or cytokine pathways (PID_IL8_CXCR2_PATHWAY, SA_MMP_CYTOKINE_CONNECTION). Helper T cell differentiation also seems affected (BIOCARTA_TH1TH2_PATHWAY), but other pathways are difficult to interpret, such as the type I diabetes gene set.

Table 5: top ten results from GSEA for DNAm age acceleration as co-variate. Shown are gene sets, their size, enrichment score (ES), normalized enrichment score (NES), nominal p-value, FDR q-value and max rank obtained during GSEA analysis.

NAME	SIZE	ES	NES	NOM p-val	FDR q-val	RANK AT MAX
REACTOME_ABORTIVE_ELONGATION_OF_HIV1_TRANSCRIPT_IN_THE_ABSENCE_OF_TAT	23	-0.628	-1.689	0.026	0.247	2759
REACTOME_PROTEIN_FOLDING	51	-0.499	-1.691	0.014	0.254	3620
REACTOME_PREFOLDIN_MEDIATED_TRANSFER_OF_SUBSTRATE_TO_CCT_TRIC	27	-0.649	-1.665	0.023	0.254	3414
REACTOME_RNA_POL_II_TRANSCRIPTION_PRE_INITIATION_AND_PROMOTER_OPENING	40	-0.590	-1.670	0.021	0.254	2759
REACTOME_RNA_POL_I_RNA_POL_III_AND_MITOCHONDRIAL_TRANSCRIPTION	115	-0.525	-1.709	0.018	0.255	4379
REACTOME_PROCESSING_OF_CAPPED_INTRON_CONTAINING_PRE_MRNA	136	-0.563	-1.675	0.046	0.255	3732
REACTOME_PROCESSING_OF_CAPPED_INTRONLESS_PRE_MRNA	23	-0.681	-1.696	0.012	0.255	4933
REACTOME_MICRORNA_MIRNA_BIOGENESIS	22	-0.602	-1.702	0.028	0.256	2759
REACTOME_FORMATION_OF_RNA_POL_II_ELONGATION_COMPLEX_	43	-0.567	-1.735	0.012	0.257	2759

The results for DNAm age acceleration are even more difficult to categorize (Table 5). The only significant gene set is related to the abortion of an HIV1 transcript (REACTOME_ABORTIVE_ELONGATION_OF_ HIV1_TRANSCRIPT_IN_THE_ABSENCE_OF_TAT).

Table 6: top five up/down regulated genes associated with DNAm age obtained with the SAM method (SAMseq function).

Gene Name	Gene ID	Score(d)	q-value(%)	Description
UPREGULATED				
GAREM1	64762	0.413	0	GRB2 associated regulator of MAPK1 subtype 1
FNIP2	57600	0.388	0	folliculin interacting protein 2
ASPHD2	57168	0.379	0	aspartate beta-hydroxylase domain containing 2
ADCY9	115	0.374	0	adenylate cyclase 9
MBP	4155	0.37	0	myelin basic protein
DOWNREGULATED				
FZD9	8326	-0.405	0	frizzled class receptor 9
PRDM13	59336	-0.387	0	PR domain 13
TRIM24	8805	-0.381	0	tripartite motif containing 24
NHLRC1	378884	-0.38	0	NHL repeat containing E3 ubiquitin protein ligase 1
HDAC2	3066	-0.378	0	histone deacetylase 2

Table 7: top five up/down regulated genes associated with DNAm age acceleration obtained with the SAM method (SAMseq function).

Gene name	Gene ID	Score(d)	q-value(%)	Description
UPREGULATED				
GAREM1	64762	0.433	0	GRB2 associated regulator of MAPK1 subtype 1
MBP	4155	0.417	0	myelin basic protein
MAPRE2	10982	0.413	0	microtubule associated protein RP/EB family member 2
SIDT2	51092	0.397	0	SID1 transmembrane family member 2
ZNF407	55628	0.395	0	zinc finger protein 407
DOWNREGULATED				
TIMM17B	10245	-0.406	0	translocase of inner mitochondrial membrane 17 homolog B (yeast)
TIMM8A	1678	-0.404	0	translocase of inner mitochondrial membrane 8 homolog A (yeast)
HDAC2	3066	-0.387	0	histone deacetylase 2
PQBP1	10084	-0.383	0	polyglutamine binding protein 1
ANAPC7	51434	-0.381	0	anaphase promoting complex subunit 7

For the SAM method, we only represented the top five up/down regulated genes (Table 6; Table 7), the whole result files being in the data output directory. Globally, the expressed genes are differently associated to DNAm age and age acceleration. Nevertheless it is interesting to note that with increasing DNAm age and acceleration, GAREM1 is upregulated, which is a regulator of MAPK1, a gene that codes for a MAP kinase involved in various cellular processes such as proliferation, differentiation and transcription regulation. Also in both cases, HDAC2 is downregulated, which codes for a histone deacetylase.

5.3. CONCLUSION

This part of the project allowed us to familiarize us with the GSEA and SAM methods, both being powerful and interesting tools for the exploration of expression datasets. While both produced significant genes and pathways associated either with DNAm age or age acceleration, the interpretation of these results is a daring task. As shown in the results, most of the correlated pathways show little in common. But, one must also admit that a specialized knowledge in colon-rectal cancer biology is needed as to estimate how the obtained results fit into known disease progression models.

The GSEA results also showed that, while a good number of gene sets had a p-value < 0.05, these where not necessarily paired with good FDR q-values, which reinforces the idea that gene selection methods only based on (uncorrected) p-values should be approached with skepticism. This also opens up the discussion concerning the q-value threshold: the GSEA team uses a threshold of 0.250, their justification being that the main objective of a GSEA analysis lies in the stipulation of hypotheses. With this in mind, the resulting significant gene sets are to be subjected to further experimental dissection and only then to be implemented in a theoretical framework for the underlying biological process. If we would have applied a stricter threshold for the FDR q-value, e.g. 0.1 or 0.05, then all our results would have been non-significant.

Finally, both DNAm age and age acceleration seem to be correlated with the regulation of GAREM1. This regulator of MAPK1 could be a potential indicator of the underlying oncological processes in the colon-rectal adenocarcinomas, since it has been known for a while that the MAPK pathway is involved in numerous cancers (Dhillon et al., 2007).

6. REFERENCES

Barba, M., Czosnek, H., Hadidi, A., 2014. Historical Perspective, Development and Applications of Next-Generation Sequencing in Plant Virology. Viruses 6, 106–136. doi:10.3390/v6010106

Dhillon, A.S., Hagan, S., Rath, O., Kolch, W., 2007. MAP kinase signalling pathways in cancer. Oncogene 26, 3279–3290. doi:10.1038/sj.onc.1210421

Horvath, S., 2015. Erratum to: DNA methylation age of human tissues and cell types. Genome Biol. 16, 96. doi:10.1186/s13059-015-0649-6

Horvath, S., 2013. DNA methylation age of human tissues and cell types. Genome Biol. 14, 3156. doi:10.1186/gb-2013-14-10-r115

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. 102, 15545–15550. doi:10.1073/pnas.0506580102

Tomczak, K., Czerwińska, P., Wiznerowicz, M., 2015. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. 19, A68–A77. doi:10.5114/wo.2014.47136

Tusher, V.G., Tibshirani, R., Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc. Natl. Acad. Sci. 98, 5116–5121. doi:10.1073/pnas.091062498

7. ANNEXES

Files annexed to this document:

Project_2016.zip: contains the 3 R scripts used for this project (*question_one.R, question_two.R* and *question_three.R*), as well as all the input data folder (*clinical_factors, mRNA_gene* and *mRNA_gene_normalized*) and finally all the output directories containing all the produced results and plots (*data_output, plots* and *GSEA_output*).

At the time of redaction of this document, the project directory is accessible through the following link: https://www.dropbox.com/s/176hlqqaucjlaoz/project_2016_rodriguesV_joel.tar.gz?dl=0

Sessioninf()o output for each script file:

question_one.R

```
R version 3.3.0 (2016-05-03)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.4 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  base

other attached packages:
[1] gridExtra_2.2.1 reshape2_1.4.1  ggplot2_2.1.0   stringr_1.0.0

loaded via a namespace (and not attached):
 [1] labeling_0.3     colorspace_1.2-6 scales_0.4.0     plyr_1.8.3
 [5] magrittr_1.5     tools_3.3.0      gtable_0.2.0     Rcpp_0.12.5
 [9] stringi_1.0-1    grid_3.3.0       methods_3.3.0    munsell_0.4.3
```

question_two.R

```
R version 3.3.0 (2016-05-03)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.4 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
```

```
 [3] LC_TIME=en_US.UTF-8       LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8   LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8      LC_NAME=C
 [9] LC_ADDRESS=C              LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] stats     graphics  grDevices utils     datasets  base

other attached packages:
[1] survival_2.39-4 gridExtra_2.2.1 plyr_1.8.3      PMCMR_4.1
[5] reshape2_1.4.1  ggplot2_2.1.0   stringr_1.0.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.5      lattice_0.20-33  grid_3.3.0       gtable_0.2.0
 [5] magrittr_1.5     scales_0.4.0     stringi_1.0-1    Matrix_1.2-6
 [9] labeling_0.3     splines_3.3.0    tools_3.3.0      munsell_0.4.3
[13] colorspace_1.2-6 methods_3.3.0
```

question_three.R

```
R version 3.3.0 (2016-05-03)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 14.04.4 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] parallel  stats4    methods   stats     graphics  grDevices utils
[8] datasets  base

other attached packages:
 [1] samr_2.0           matrixStats_0.50.2   impute_1.46.0
 [4] org.Hs.eg.db_3.3.0 AnnotationDbi_1.34.2 IRanges_2.6.0
 [7] S4Vectors_0.10.0   Biobase_2.32.0       BiocGenerics_0.18.0
[10] survival_2.39-4    gridExtra_2.2.1      plyr_1.8.3
[13] PMCMR_4.1          reshape2_1.4.1       ggplot2_2.1.0
[16] stringr_1.0.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.5      magrittr_1.5     splines_3.3.0    munsell_0.4.3
 [5] colorspace_1.2-6 lattice_0.20-33  tools_3.3.0      grid_3.3.0
 [9] gtable_0.2.0     DBI_0.4-1        Matrix_1.2-6     RSQLite_1.0.0
[13] stringi_1.0-1    scales_0.4.0
```