


Project report

DNAm age in the head and neck squamous cell carcinoma



Author François Kroll

Professor Vincent Detours

Academic year 2017-2018

BIOL-F-423 – Analysis of Functional Genomic Data

Master's in Bioinformatics and Modelling, Université

Libre de Bruxelles (ULB), Belgium



UNIVERSITÉ
LIBRE
DE BRUXELLES

Introduction

In this work, we set up to explore the potential relationships between the age of head and neck squamous cell (HNSC) tumours, as measured by DNA methylation (DNAm), and different variables, such as patients' clinical annotations, survival and gene expression.

Common material and methods

The threshold for significance of the p-value is considered to be 0.05 throughout this work.

All data, except for Q3.1 which uses the software GSEA¹, were analysed and plotted with RStudio v1.0.143, functioning with R language v3.4.0². If specific packages were used, they are mentioned and referenced in the *Material and methods* section of each question.

Except from figures from GSEA, all plots were also created with R². In the figures, an asterisk (*) indicates a p-value below 0.05, two asterisks (**) indicate a p-value below 0.01.

All input data were downloaded from GDAC FireHose³, except for the DNAm ages which were received from Mr. Detours.

Q1: DNAm age vs. chronological age

This section (question 1) will first investigate possible correlations between the DNAm age of tumorous and normal tissues with the patient' chronological age. It will then examine the correlation between the age of tumorous tissues and the age of normal tissues. Finally, it will test two possible definitions of "tumour age acceleration".

Material and methods

The input consisted of a file downloaded from *The Cancer Genome Atlas* (TCGA)³: the main file containing patients' clinical annotations (*HNSC.clin.merged.picked.txt*, from *Clinical_Pick_Tier1*), and the file sent by Mr. Detours containing DNAm ages (*HNSC-8.Rda*).

The DNAm ages file contained data for 400 samples. Among them, 349 were DNAm ages of primary tumours (tissue code 01), 50 were DNAm ages of normal tissues (tissue code 11), and 1 was the DNAm age of a metastase (tissue code 06). The samples first originated from 349 distinct patients. The clinical annotation file contained annotations for 528 patients.

Briefly, from the clinical annotation file, only the patients who had a DNAm age for their primary tumour were kept. That represented 349 patients. One patient (patient code *a4ca*) did not have a chronological age in the clinical file, and was removed from both files. From the DNAm ages file, the DNAm age of a metastase tissue (tissue code 06) was removed, and the file was split into

the DNAm ages of primary tumours and the DNAm ages of normal tissues, keeping track of the patient from who each measurement originated. Finally, these three files (normal tissues DNAm ages, primary tumours DNAm ages, clinical data) were merged into one single data frame by matching the patient codes. The final data frame thus has 348 rows, which each represents a patient; and the columns include their chronological ages (all patients have a reported chronological age), the DNAm ages of their tumours (all patients have a measured DNAm tumour age), the DNAm ages of normal tissues (only 50 patients have a measured DNAm normal tissue age), and clinical annotations.

Correlations between tumour ages and chronological ages were first visually assessed by a scatter plot. Then, the Spearman rank-order correlation was computed. A p-value was also computed by asymptotic *t* approximation, using R commands like:

```
cor.test (Clinical$tumour_age, Clinical$normaltissue_age, use = 'pairwise.complete.obs',  
          method = 'spearman')
```

where *use = 'pairwise.complete.obs'* tells R to only use all complete pairs of observations for the two variables being compared. This is important when comparing variables which include missing data (NAs), such as the ages of the normal tissues.

The Pearson product moment correlation would probably have been similarly adequate in this setting, given the relatively normal distributions of the chronological, tumour DNAm and normal tissue DNAm ages (see Q2.1). However, Spearman's correlation has the advantage of detecting monotonic relationships, and not only linear like Pearson's. All results were subsequently confirmed with a Pearson's correlation: results were always similar (not shown) and never changed the overall conclusions.

1 – Tumour DNAm age vs. chronological age

Results

The tumour DNAm age vs. chronological age scatter plot (**Figure 1A**) does not show a strong correlation between these two variables. This is confirmed by the low correlation coefficient ($r = 0.21$), which indicates a weak correlation. The p-value is very low (p-value = 6.552×10^{-5}), indicating that the correlation, despite being weak, is highly significant. The p-value benefits from the large sample size ($N = 348$), which allows the correlation to be detected as significant despite the low correlation coefficient.

Discussion

Tumour DNAm age is only very weakly positively correlated to the patient's chronological age. This conclusion is in accordance with Horvath's results⁴ (Additional file 13J): he found a correlation of 0.13 (vs 0.21 for us) and a p-value of 0.022 (vs. 6.552×10^{-5} for us) for tumour DNAm age vs. patient chronological age.

2 – Normal tissue DNAm age vs. chronological age

Each normal tissue is a sample of non-cancerous tissue taken from the same anatomic site than the primary tumour⁵.

Results

The normal tissue DNAm age vs. chronological age scatter plot (**Figure 1B**) shows a clear positive correlation between the two variables: when the patient's age increases, the DNAm age of the normal tissues tends to increase too. This is confirmed by the Spearman's correlation coefficient ($r = 0.52$) which indicates a moderate correlation (Pearson's correlation coefficient is 0.67, which would indicate a strong correlation). Furthermore, this correlation is highly significant (p-value = 0.0001144).

Discussion

There is a clear positive correlation between normal tissue DNAm age and the patient's chronological age. This relationship is expected and is known as the DNA methylation clock hypothesis⁴. Again, these results are in accordance with Horvath's conclusion (Figure 1 of the article): he found a correlation of 0.73 (vs. 0.52 for us) and a p-value of 1.8×10^{-9} (vs. 0.0001144 for us) for normal head and neck tissue DNAm age vs. patient chronological age.

The fact that the patient's chronological age does not correlate with the DNAm age of his/her tumour (Q1.2, previous section) but does correlate with the DNAm age of his/her normal tissue is interesting. HNSC cancer tissues somehow do not follow the patient's overall aging process.

3 – Tumour DNAm age vs. normal tissue DNAm age

Results

Looking at the scatter plot (**Figure 1C**), there does not seem to exist a strong correlation between the DNAm age of the tumour and the DNAm age of the normal tissue of the same patient. The correlation coefficient points in the same direction ($r = 0.23$), indicating a weak correlation. Nevertheless, this weak correlation is not significant (p-value = 0.1078). The p-value probably suffers here from the low sample size ($N = 50$).

Conclusion

Horvath did not test (or did not include) these results in his paper. Nevertheless, as the two previous sections confirm his results, it is likely that he would have reached the same conclusion. Indeed, if the chronological age correlates with the normal tissue DNAm age but does not correlate with the tumour DNAm age, it is unlikely that the two DNAm ages correlate together.

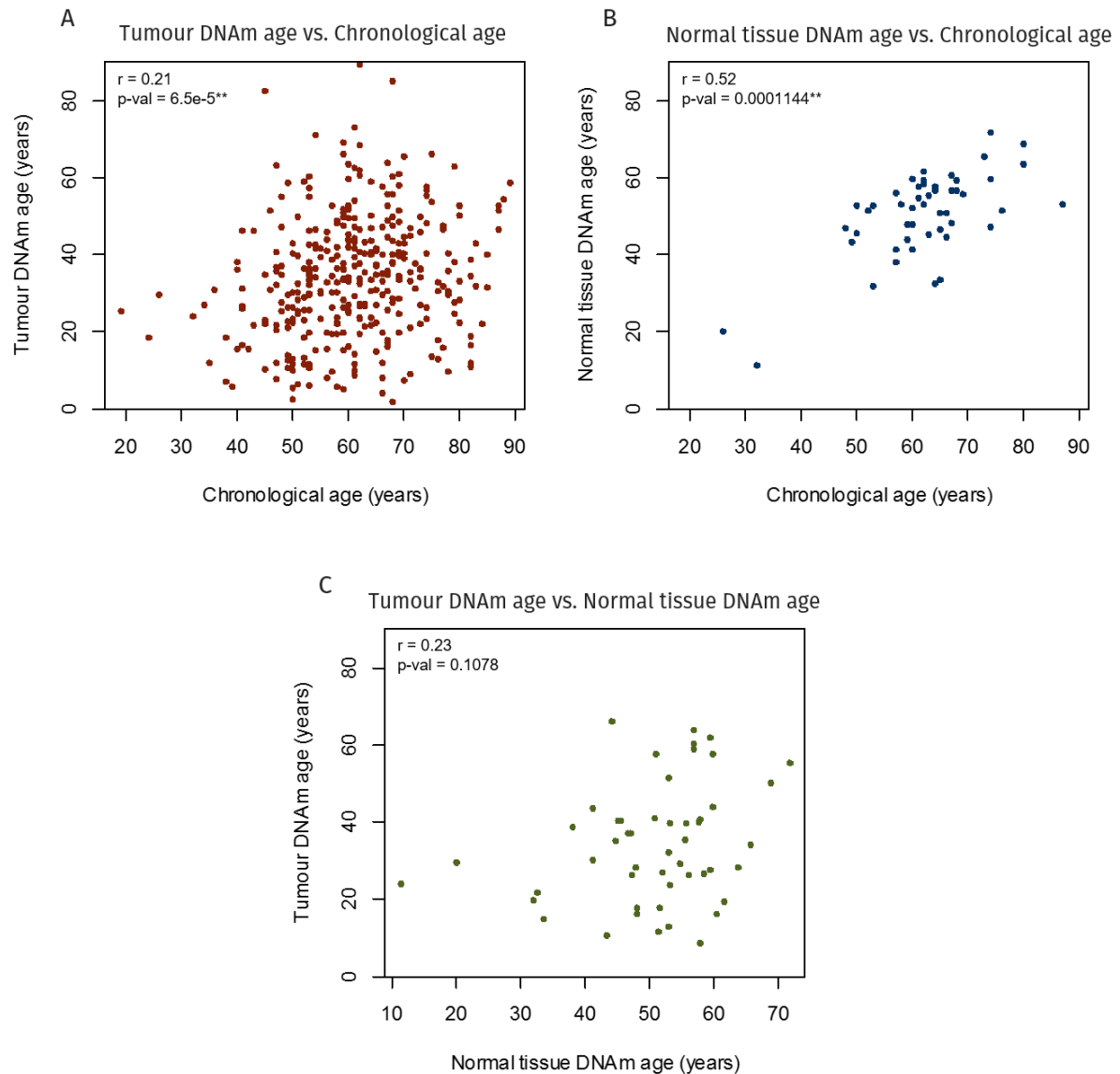


Figure 1 | Questions 1.1–1.3. (A) Question 1.1: tumour DNAm age vs. patient chronological age scatter plot. $N = 348$ patients. **(B)** Question 1.2: normal tissue DNAm age vs. patient chronological age scatter plot. $N = 50$ patients. **(C)** Question 1.3: tumour DNAm age vs. normal tissue DNAm age scatter plot. $N = 50$ patients. Correlation coefficients and p-values are included in each plot.

4 – Tumour age accelerations

The tumour age acceleration represents the ratio between the tumour DNAm age and the normal age. The normal age can be defined either as the normal tissue DNAm age (definition 1 of the tumour age acceleration) or the patient's chronological age (definition 2 of the tumour age acceleration). It can intuitively be interpreted as how much younger (if the acceleration is < 1) or older (if the acceleration is > 1) the tumour is in terms of DNAm age compared to the normal tissue or the patient's chronological age, respectively for definition 1 and 2.

Stated in more mathematical terms:

$$\text{Definition 1} \quad \text{tumour acceleration}_1 = \frac{\text{tumour DNAm age}}{\text{normal tissue DNAm age}} \quad (1)$$

$$\text{Definition 2} \quad \text{tumour acceleration}_2 = \frac{\text{tumour DNAm age}}{\text{chronological age}} \quad (2)$$

Results

Tumour age accelerations were computed for all the samples. As all patients have a tumour DNAm age and a chronological age, there are 348 tumour accelerations₂ (one per patient). However, only 50 patients among the 348 have a normal tissue DNAm age, there are thus only 50 tumour accelerations₁.

As shown by the tumour age acceleration₁ vs. tumour age acceleration₂ scatter plot (**Figure 2A**), these two definitions overlap, in the sense that they are very strongly correlated ($r = 0.93$, $p\text{-value} \approx 0$). Using definition 1 or definition 2 should thus be essentially equivalent.

To further explore how close these definitions are, the Kernel density plots for each definition of the tumour acceleration were plotted and superimposed (**Figure 2B**). The plot clearly shows that both tumour accelerations are similarly distributed.

Finally, it would be possible to compare these distributions in more statistical terms with the two-sample Kolmogorov–Smirnov test. The $p\text{-value}$ is 0.01978. Being lower than the conventional 0.05 threshold, it would indicate that the two distributions are unlikely to be equivalent. However, given the very high correlation and the density plots, it can safely be assumed that these two definitions are essentially similar for our applications. Given the higher number of samples for definition 2 of tumour age acceleration (50 for definition 1 vs. 348 for definition 2), this definition will be preferred throughout this work.

Conclusion

Definition 2 of the tumour age acceleration will be used throughout this work. This definition of tumour age acceleration is also close (in the sense that it uses the same two variables) to the definition used by Horvath, which is *tumour DNAm age – chronological age*.

As regard with Horvath's results⁴, he mentions an average age acceleration across all cancers of 36.2 years, and the average for HNSC tumours specifically is around 25 years (the exact result is not included) (Additional file 13B). When Horvath's definition of age acceleration is used, we reach a strikingly similar result, but opposite (average age acceleration = -26.72 years). It seems very unlikely to obtain similar but opposite results. An explanation might be that Horvath used the absolute value of this difference, although this does not seem to be literally mentioned anywhere in the article. If that was the case, it may be a questionable choice as the sign appears to be crucial here: this negative value represents a tumour age "deceleration", i.e. the tumour appears younger than the patient, rather than an age "acceleration", i.e. the tumour appears older than the patient.

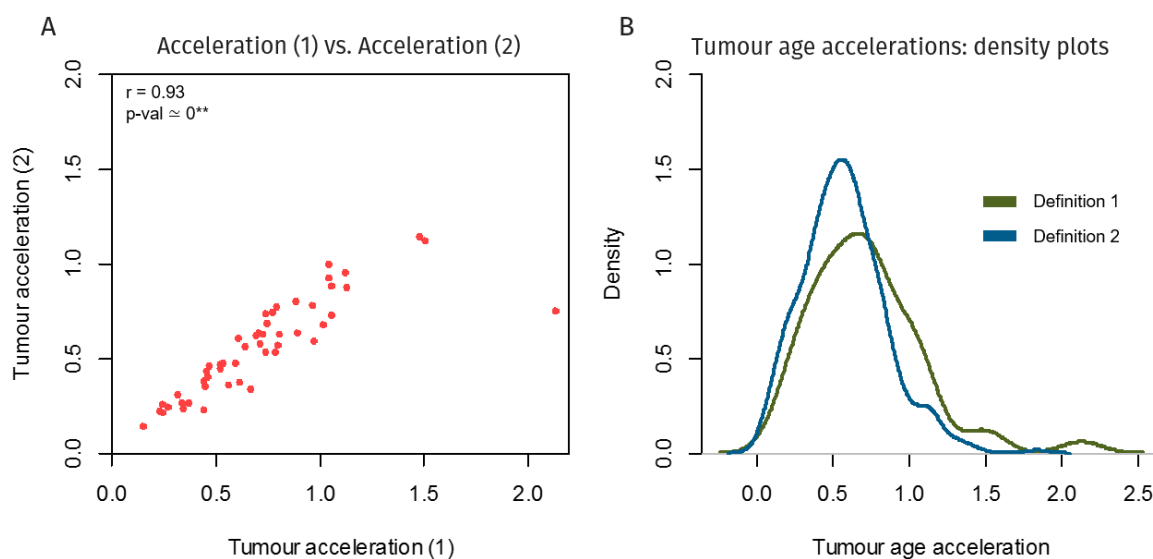


Figure 2 | Question 1.4. **(A)** Definition 1 of tumour age acceleration vs. definition 2 of tumour age acceleration scatter plot. **(B)** Kernel density plots showing the distribution of the tumour age accelerations whether definition 1 (green) or definition 2 (blue) is used.

Q2.1: Tumour DNAm age vs. clinical variables

This section will investigate possible relationships between the tumour DNAm age and the tumour age acceleration with clinical variables. For each clinical variable, first its relationship with the tumour DNAm age will be assessed, then its relationship with the tumour age acceleration (definition 2).

Material and methods

The input consisted of the same files as Q1 (see the *Material and methods* section of Q1), with the additional file *All_CDEs.txt* (also from *Clinical_Pick_Tier1*³).

All clinical variables in the file *HNSC.clin.merged.picked.txt* were studied, and a series of additional clinical variables were transferred from *All_CDEs.txt*. They were selected if there could reasonably be a relationship with a DNAm tumour age and if the data were not predominantly missing values (NAs). The number of observations is indicated in the legend of each figures. Some clinical variables from *All_CDEs.txt* that did not produce significant results were not included.

The selection of the appropriate statistical test is primarily dictated by whether the distribution of the data can be considered normal or not. For this purpose, the distribution of all the continuous variables tested was checked for normality. This verification consisted in a visual assessment of the density plot and of the Q-Q plot. The Q-Q plot is a graphical tool to help in assessing whether data follow a theoretical distribution (e.g. **Figure 3BD**), in our case the normal distribution. It is a form of scatter plot where the Y coordinate of each point is a quantile of our data and the X coordinate is the theoretical quantile of the normal distribution. If the points approximately follow a straight line (it is represented as a grey line in **Figure 3BD**), it means our data set probably came from a theoretical normal distribution, and thus that it is safe to make the normality assumption⁶.

The Shapiro-Wilk test could be used to formally test the normality of a distribution. However, it is mostly inadequate in our setting due to the large sample size. Indeed, when the sample size is large, any small deviation from the normal will cause the Shapiro-Wilk test to return a significantly low p-value, which indicates that the distribution should not be considered as normal. Moreover, formal normality check is mostly unnecessary when the sample size is large (approximately $N > 30$) due to the “central limit theorem”, which states that even if a population is non-normally distributed, the distribution of the “sample means” will be normally distributed if the sample size is large⁷. For this reason, most statistical tests assuming normality, including the Welch’s t-test and the one-way ANOVA (see hereafter) remain valid for moderate deviations from normality when the sample size is large.

Overall, when N is large and when the only goal is to decide if tests assuming normality of the distribution can be used, formally testing the normal distribution with a Shapiro-Wilk test is often unnecessary and the Q-Q plot should be relied upon. This is the methodology used in this work.

Plots and statistical tests also differ according to the nature of the clinical variable.

If the clinical variable is continuous (e.g. number of cigarette packs smoked per year), a simple scatter plot is used to display the relationship with tumour DNAm age or tumour age acceleration. If the distributions of each of the two variables can be considered normal, they are statistically compared with the Welch's t-test. It is a variant of the Student's t-test which is more reliable if the variances are unequal and/or if sample sizes are unequal, which is the case in most of our comparisons. If the distributions cannot be considered normal, the Mann-Whitney *U* test (also called Wilcoxon rank sum test) should be used.

If the clinical variable is categorical or binary, the relationship is displayed with a boxplot. The upper limit of the box represents the first quartile, the lower limit the third quartile, and the black line inside the box the median value. The upper whisker reaches the maximum value, and the lower whisker reaches the minimum value. Outliers are represented as additional points. In this case, the two variables are statistically compared with the one-way ANOVA (analysis of variance) if the distribution of the tumour DNAm age or the tumour age acceleration can be considered normal. If the distributions cannot be considered normal, the Kruskal-Wallis one-way analysis of variance should be used.

Normal distributions

Tumour DNAm age

As explained previously, the distribution of the tumour DNAm age should first be assessed for normality. The density plot (**Figure 3A**) and the Q-Q plot (**Figure 3B**) both clearly indicate that the distribution of the tumour DNAm age is normal enough to perform tests that assume normality of the distribution. The Welch's t-test and the one-way ANOVA will thus be used in this section for binary and categorical (more than 2 categories) variables, respectively.

Tumour age acceleration

Distribution of the tumour age acceleration (definition 2) should also be checked for normality. The density plot (**Figure 3C**) and the Q-Q plot (**Figure 3D**) also indicate that the distribution is normal enough to perform the Welch's t-test and the one-way ANOVA.

Conclusion

We can safely assume that the tumour DNAm ages and the tumour age accelerations follow normal distributions. Hence, parametric tests, i.e. tests which assume that the distribution is normal, can be used. Namely, that is the Welch's t-test or the one-way ANOVA in our case.

Nonetheless, given the simplicity and rapidity of running the statistical tests, results were subsequently confirmed with the corresponding non-parametric test, i.e. which does not assume distribution normality. Namely, that is the Mann-Whitney U test or the Kruskal-Wallis one-way analysis of variance in our setting. These results are not included here. The p-values were always similar and never changed the overall conclusions.

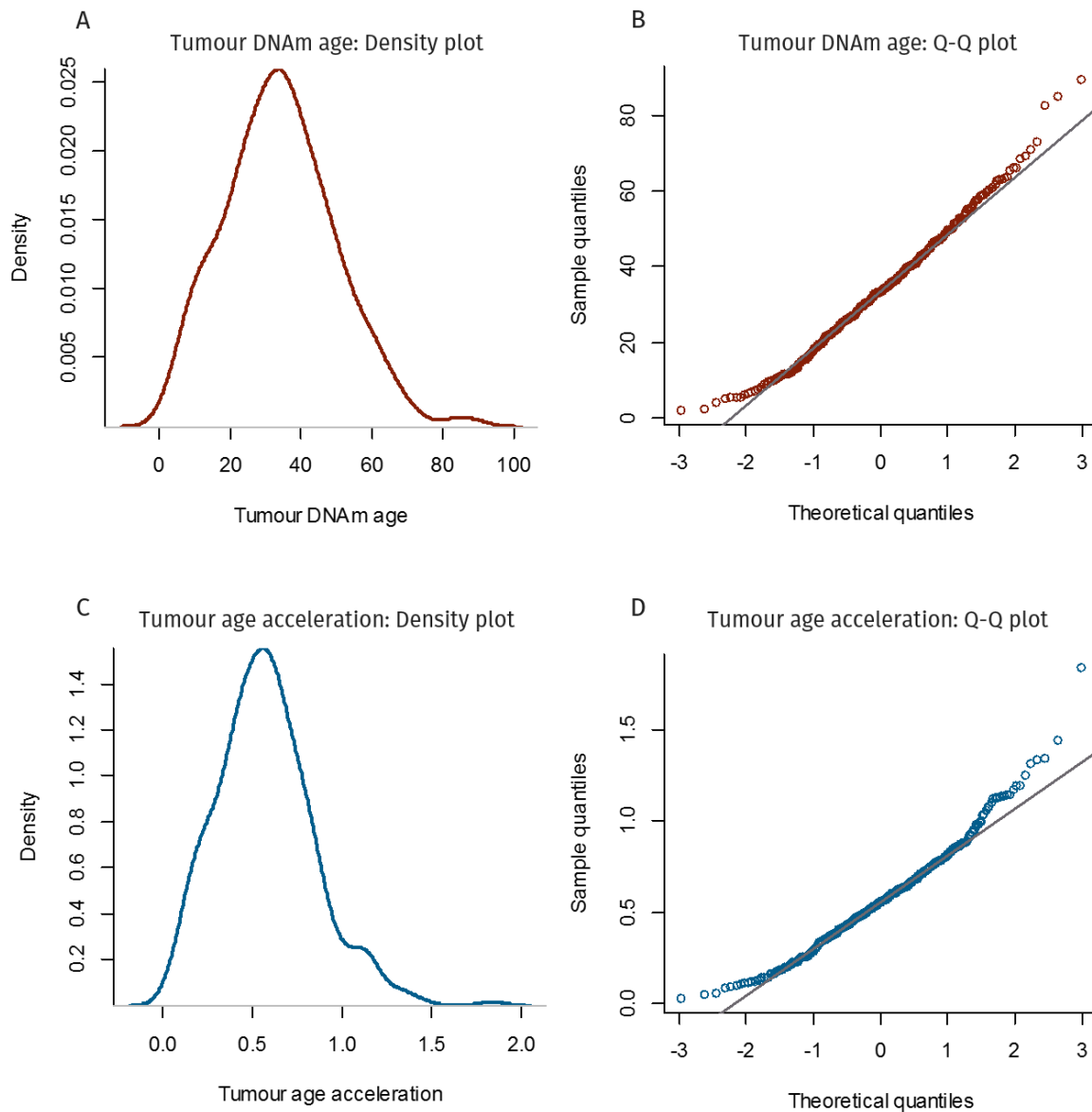


Figure 3 | Assessing the distribution normality assumption for the tumour DNAm age and the tumour age acceleration. **(A)** Kernel density plot of the tumour DNAm age. **(B)** Q-Q plot of the tumour DNAm age. **(C)** Kernel density plot of the tumour age acceleration. **(D)** Q-Q plot of the tumour age acceleration. $N = 348$ patients for all four plots.

Vital status

The vital status is a binary variable. It is either 1 if the patient is dead, or 0 if the patient is alive. There are no missing data for this variable (154 dead vs. 194 alive). This variable will also be used in the survival analysis (see Q2.2).

Tumour DNAm age

The boxplot (**Figure 4A**) does not seem to show any significant difference in the tumour DNAm age whether the patient is dead or alive. This is confirmed by the results of the Welch's t-test (p-value = 0.9158).

Tumour age acceleration

Based on the boxplot (**Figure 4B**), there does not seem to exist any significant difference in the tumour age acceleration whether the patient is dead or alive. This is confirmed by the results of the Welch's t-test (p-value = 0.2403).

Conclusion

The age of the tumour, whether it is measured as absolute DNAm age or as a ratio of the patient's chronological age, does not seem to affect the patient's vital status. At first approximation, this may indicate that tumour age does not influence prognosis, although this question will be more formally addressed with the survival analysis.

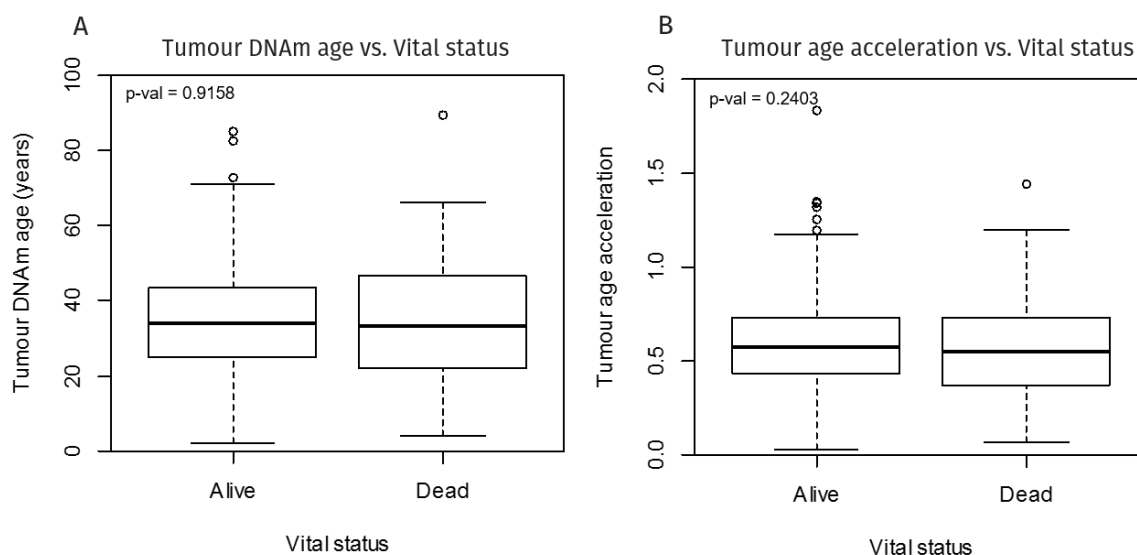


Figure 4 | Tumour age vs. vital status. **(A)** Tumour DNAm age vs. vital status boxplot. **(B)** Tumour age acceleration vs. vital status boxplot. The p-value computed by the Welch's t-test is included in each plot. N = 348 patients for both plots.

Gender

The gender is a binary variable: male or female. There are no missing data for this variable (92 female vs 256 male).

Tumour DNAm age

At first sight at the boxplot (**Figure 5A**), there does not seem to exist any meaningful difference in the tumours' DNAm ages in the female population or the male population. This is confirmed by the results of the Welch's t-test (p-value = 0.1602).

Tumour age acceleration

The boxplot (**Figure 5B**) does not indicate any difference in tumour age accelerations in the female population or the male population. This is confirmed by the results of the Welch's t-test (p-value = 0.7741).

Conclusion

HNSC tumours' ages, as measured by DNA methylation, do not significantly differ according to gender. In other words, female and male patients do not have significantly “younger” or “older” HNSC tumours.

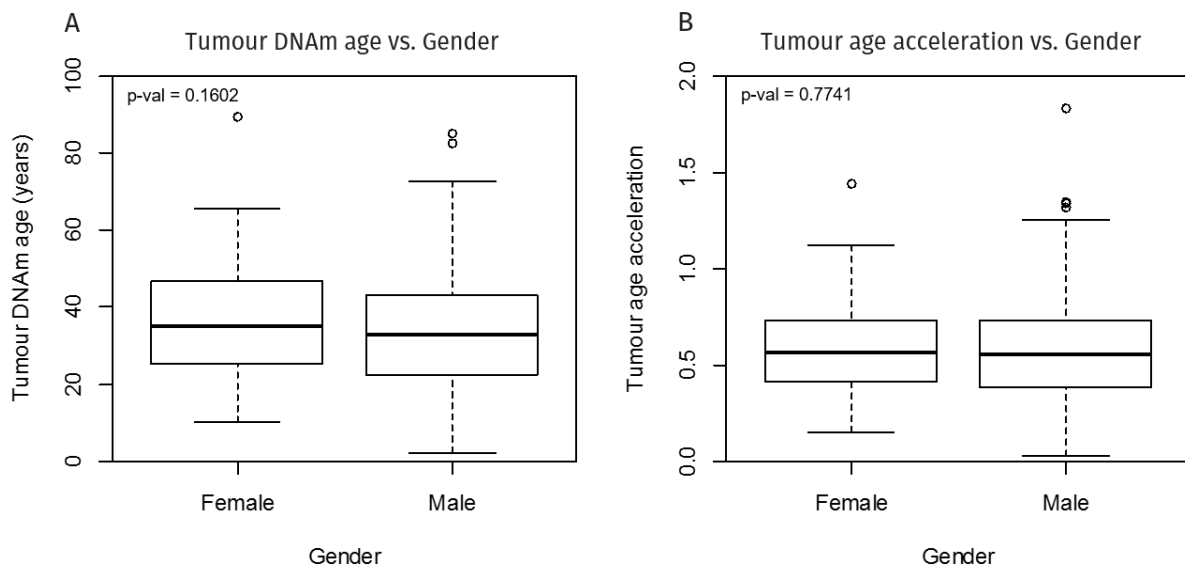


Figure 5 | Tumour age vs. patient gender. (A) Tumour DNAm age vs. patient gender. (B) Tumour age acceleration vs. patient gender. The p-value computed by the Welch's t-test is included in each plot. N = 348 for both plots.

Radiation therapy

Radiation therapy is a binary variable: *yes*, if the patient was treated with radiation therapy; or *no*, if the patient was not treated with radiation therapy. It is not entirely clear if the tumour DNAm age measurements were performed before or after the radiation therapy (in the *yes* case). If it is before, this variable is of little use, as we would merely measure the relationship between tumour age and the likelihood that the doctor prescribes radiation therapy. If it is after, as expected, the variable is much more interesting and would measure if radiations have an effect on the DNAm age of the tumour. How long is the period between the radiations and the DNAm measurements should probably be considered too. It will be assumed in the interpretation of these results that the tumour DNAm age was measured after radiation therapy, and approximately after the same period of time for all patients. Overall, these results should be taken with caution as it is not fully clear what they represent.

Among the 348 patients, 189 were treated with radiation therapy, 111 were not, and 48 were not reported (NAs). It would probably be safe to assume that these 48 patients did not receive radiation therapy, and could thus be considered as *no*. However, this is unlikely to change the overall conclusion (see hereafter).

Tumour DNAm age

The boxplot (**Figure 6A**) does not indicate any relationship between the DNAm age of the tumour and whether the patient was treated with radiation therapy or not. Welch's t-test confirms this (p-value = 0.8589).

Tumour age acceleration

The boxplot for the tumour age acceleration (**Figure 6B**) does not any indicate any meaningful difference. This is confirmed by Welch's t-test (p-value = 0.5368).

Conclusion

Radiation therapy does not have any striking effect of the DNAm age of the HNSC tumour. These results should be taken with caution as they assume that DNAm measurements were performed after the radiation therapy, which is not explicitly mentioned.

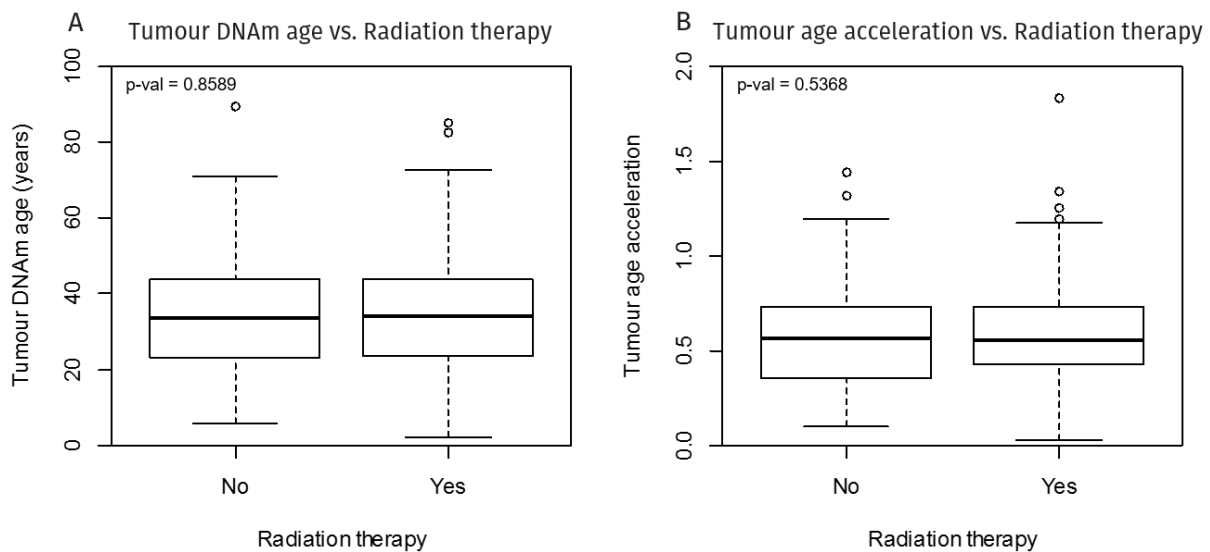


Figure 6 | Tumour age vs. whether the patient received radiation therapy (Yes) or not (No). **(A)** Tumour DNAm age vs. radiation therapy. **(B)** Tumour age acceleration vs. radiation therapy. The p-value computed by the Welch's t-test is included in each plot. N = 300 patients for both plots.

Pathologic stages and metastasis

Four clinical variables which describe the cancer's pathologic stage will be analysed here. The pathologic stages are based on the TNM classification system⁸.

First, the primary tumour pathologic stage, which is abbreviated as the T pathologic stage. As its name suggests, it describes the primary tumour, i.e. the initial tumour located in some tissue of the head or neck in our case, for instance in the oral cavity. Without entering into details, it is mainly based on the size of the primary tumour and on which anatomical sites or tissues are invaded. Possible stages, in the order of progression of the cancer, are: TX (i.e. the primary tumour cannot be assessed, 27 patients); T0 (no evidence of primary tumour, 0 patient); T1 (32 patients); T2 (81 patients); T3 (73 patients); T4 (5 patients); T4a (112 patients); T4b (4 patients). The T pathologic stage of 14 patients was not reported (NAs).

Second, the regional lymph nodes pathologic stage, abbreviated as the N pathologic stage. It describes if metastases are present in neighbouring lymph nodes and how big they are if it is the case. Possible stages, in the order of severity, are: NX (i.e. the regional nodes cannot be assessed, 50 patients); N0 (no evidence of metastasis in lymph nodes, 116 patient); N1 (42 patients); N2 (6 patients); N2a (5 patients); N2b (73 patients); N2c (35 patients); N3 (6 patients). The N pathologic stage of 15 patients was not reported (NAs).

A third variable, which directly relates to the N pathologic stage, is the number of lymph nodes affected. This information is reported for 274 patients out of 348. Among these 274 patients, 110 have zero lymph node affected. This is a continuous variable, which will be analysed using a scatter plot and Spearman's correlation.

The TNM classification system includes a third pathologic stage: the distant metastasis pathologic stage, abbreviated as the M pathologic stage. This data is reported for some patients, but only one patient had a metastasis reported (i.e. M1 stage). All the other patients (347 patients) were either MX (metastasis cannot be assessed, 43 patients); M0 (no distant metastasis, 110 patients); or not reported (i.e. NA, 194 patients). For want of more M1 patients, this variable was not analysed. The N pathologic stage and the number of lymph nodes should probably be sufficient here to assess if the tumour DNAm age affects its propensity to form metastases.

Finally, the overall pathologic stage. It is a generic pathologic stage which takes into account the three stages of the TNM system. Possible stages, in the order of severity of the disease, are: stage I (20 patients); stage II (47 patients); stage III (50 patients); stage IVa (174 patients); stage IVb (10 patients); and stage IVc (1 patient). The pathologic stage of 46 patients was not reported. The single patient with a stage IVc tumour was not included in the analysis.

All three pathologic stages (T pathologic stage, N pathologic stage, overall pathologic stage) are categorical variable (more than 2 categories). Their relationship with the tumour age will thus be formally tested with the one-way ANOVA.

Tumour DNAm age

At first assessment, the boxplots (**Figure 8ABC**) do not indicate any clear relationship between the tumour DNAm age and the pathologic stages. Only patients in the N2 stage seem to have slightly “younger” tumours, but the sample size of this population is very small (6 N2 patients).

The lack of significant differences is confirmed by the one-way ANOVAs (p-value = 0.664 for the T stage; p-value = 0.632 for the N stage; p-value = 0.898 for the overall pathologic stage).

The correlation between the tumour DNAm age and the number of lymph nodes is statistically significant but very weakly negative to negligible (**Figure 7A**, $r = -0.16$, p-value = 0.008392).

Tumour age acceleration

The boxplots (**Figure 9ABC**) do not seem to indicate any striking relationship between the tumour age acceleration (definition 2) and the pathologic stages.

This is confirmed by the one-way ANOVAs (p-value = 0.739 for the T stage; p-value = 0.832 for the N stage; p-value = 0.979 for the overall pathologic stage).

Similarly to DNAm age, the correlation between the tumour age acceleration and the number of lymph nodes is statistically significant but negligible (**Figure 7B**, $r = -0.14$, p-value = 0.02068).

Conclusion

The tumour DNAm age or age acceleration does not seem to affect its pathologic stage. Stated differently, an “older” or “younger” tumour (in terms of DNAm) does not seem to be more or less aggressive. The lack of relationship between the tumour’s age and the N stage also indicate that there is probably no correlation between the M (metastasis) stage and the tumour’s age, even if the sample included more patients with metastasis (M1 stage). Indeed, it may probably be

assumed that if the tumour's age significantly affected its propensity to form metastasis, it would have been detected at the level of the regional lymph nodes metastasis.

Similarly to the vital status, this overall lack of relationship with the pathologic stage also seems to indicate that the tumour's DNAm age does not affect prognosis. Nonetheless, this question will be more formally addressed in the survival data analysis.

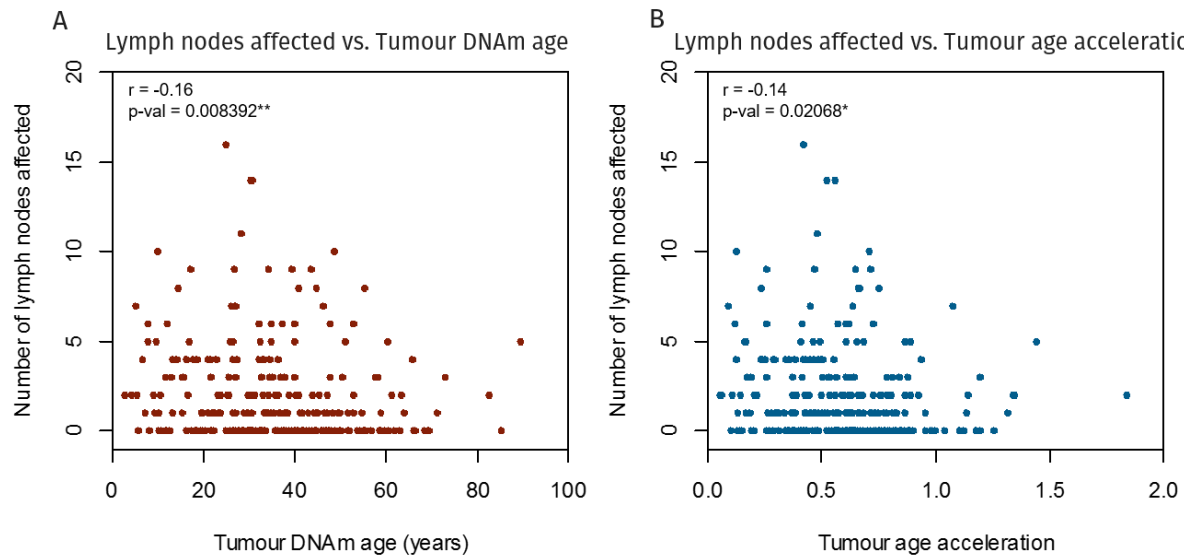


Figure 7 | Number of lymph nodes invaded by the tumour vs. its age. **(A)** Number of lymph nodes vs. tumour DNAm age. **(B)** Number of lymph nodes vs. tumour age acceleration. The Spearman's correlation coefficient and its p-value are included in each plot. $N = 274$ for both plots. One patient with 42 nodes affected was omitted in both plots for graphical reasons (273 patients actually plotted).

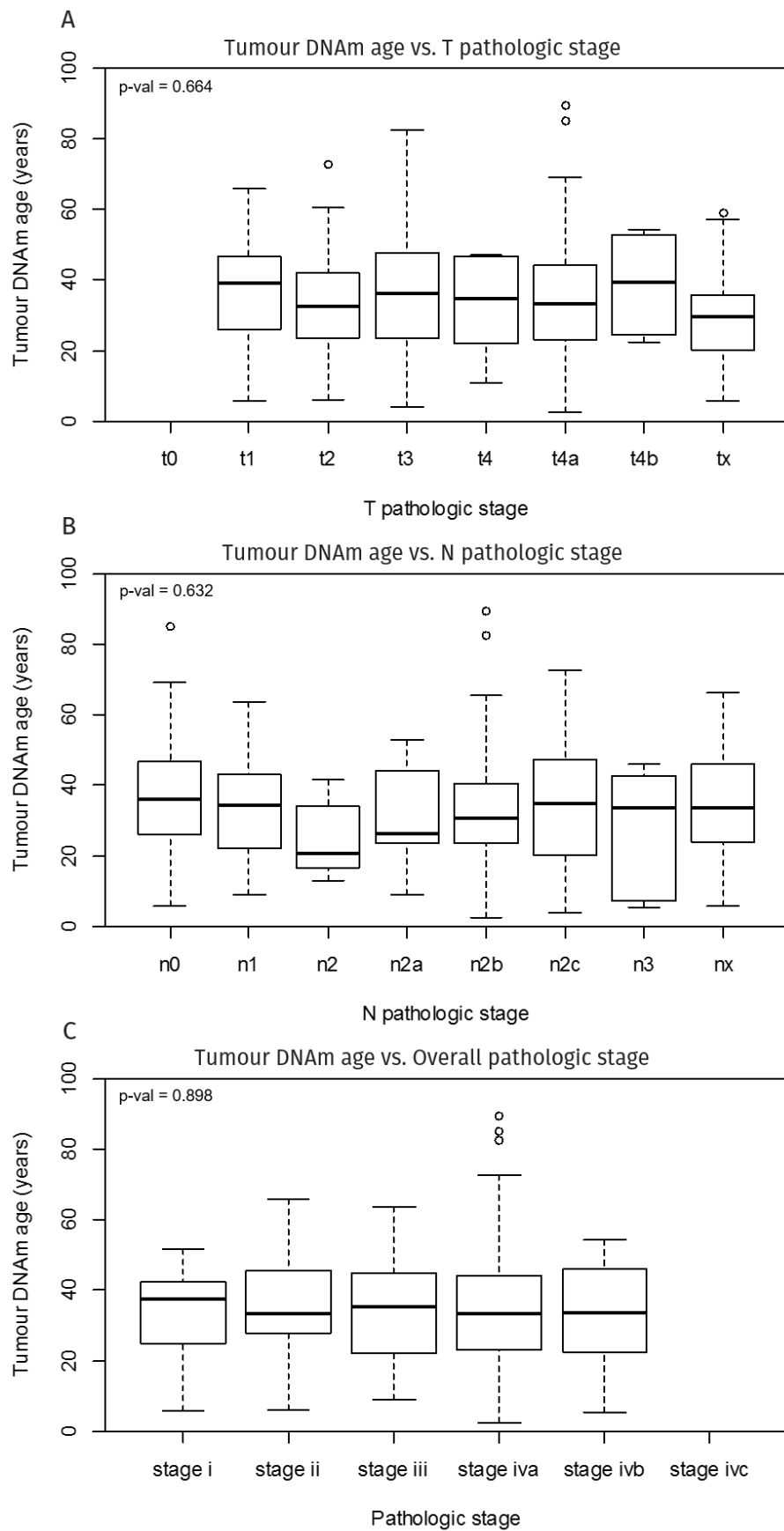


Figure 8 | Tumour DNAm age vs. pathologic stages. **(A)** vs. primary tumour pathologic stage (N = 334 patients). **(B)** vs. regional lymph nodes pathologic stage (N = 333 patients). **(C)** vs. overall pathologic stage (N = 302 patients). The p-value computed by the one-way ANOVA is included in each plot.

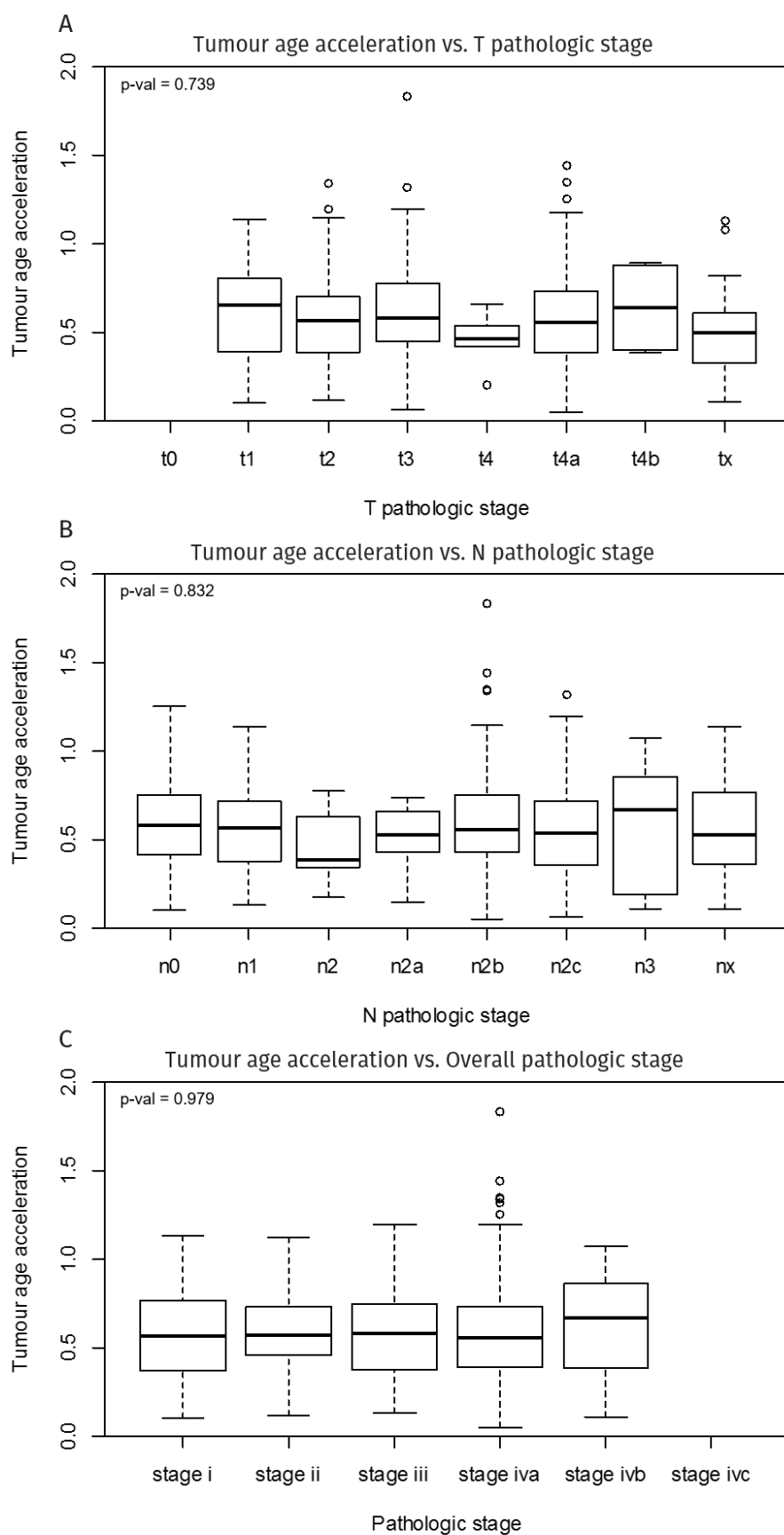


Figure 9 | Tumour age acceleration vs. pathologic stages. **(A)** vs. primary tumour pathologic stage (N = 334 patients). **(B)** vs. regional lymph nodes pathologic stage (N = 333 patients). **(C)** vs. overall pathologic stage (N = 302 patients). The p-value computed by the one-way ANOVA is included in each plot.

Race

Race is an arbitrary classification of patients based on physical attributes, history, nationality and geographic distribution. Possible races include: American Indian or Alaska Native (2 patients); Asian (8 patients); Black or African American (36 patients); White (294 patients). The race of 8 patients was not reported. It is a categorical variable and will thus be analysed like the pathologic stages (see here before).

Tumour DNAm age

The boxplot (**Figure 10A**) does not show any notable differences in the tumour DNAm age in each population. This is confirmed by the one-way ANOVA (p-value = 0.288).

Tumour age acceleration

The boxplot (**Figure 10B**) does not seem to indicate any significant difference in the tumour age acceleration in each population. This is confirmed by the one-way ANOVA (p-value = 0.228).

Conclusion

The patient's race does not seem to affect his/her tumour's DNAm age. Asian people, especially Chinese⁹, and Natives of the Arctic region¹⁰, which may be associated with the native population in our dataset, are more at risk for nasopharyngeal carcinoma (one type of HNSC tumour). Although this should be taken with extreme caution due to the very small sample sizes (8 Asian patients, 2 Native patients), our results may indicate that this risk factor is not reflected in the age of the tumour.

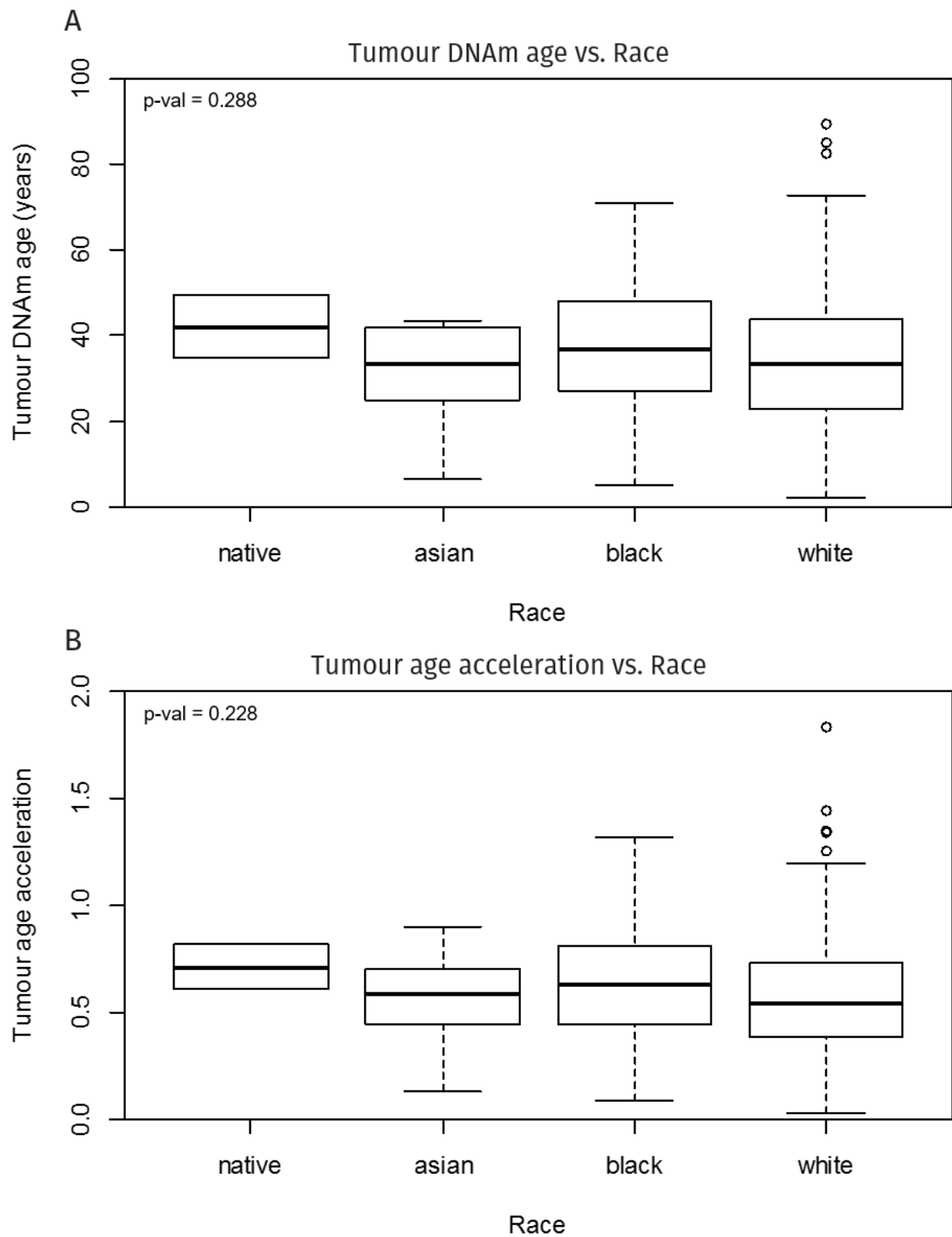


Figure 10 | Tumour age vs. race. **(A)** Tumour DNAm age vs. race. **(B)** Tumour age acceleration vs. race. The p-value computed by the one-way ANOVA is included in each plot. N = 340 patients.

Smoking behaviour

Three different clinical variables related to the patient's smoking behaviour are discussed here.

First, the year of smoking onset. It represents the date (the year) at which the patient started smoking. It is not usable at such, but by subtracting this value from the date of initial pathologic diagnosis, we can obtain an amount of years representing how long the patient has been smoking when he/her was diagnosed for HNSC cancer. We have these data for 186 patients out of 348. This variable is continuous, and will thus be analysed with a scatter plot and Spearman's correlation.

Second, the number of cigarette packs smoked per year. This data is available for 192 patients out of 348. This is a continuous variable, it will thus be analysed with a scatter plot and the Spearman's correlation.

Third, the smoker vs non-smoker status. This variable is not present in the original file. However, it can be generated from the *number of cigarette packs smoked per year* variable. It was generated by replacing available data by "yes", and by assuming that non-available (NAs) data were "no", i.e. non-smokers. These data could also have been generated from the year of smoking onset, but the former was preferred because there were more available data (192 vs 186), thus potentially including more smokers. Overall, there could be some smokers in the non-smoking group, so these results should be taken with caution. The smoker status logically includes data for all 348 patients. It is a binary variable, and will be thus analysed like, for instance, the vital status, i.e. using a boxplot and the Welch's t-test.

Tumour DNAm age

The scatter plot of the tumour DNAm age vs. the number of years smoking (**Figure 11A**) may suggest a weak to moderate correlation between the two variables. This is confirmed by the correlation coefficient and the p-value: the number of years smoking weakly positively correlates with the tumour DNAm age ($r = 0.34$), and this correlation is highly significant ($p\text{-value} = 2.139 \times 10^{-6}$).

The number of cigarette packs smoked per year does not correlate with the tumour DNAm age (**Figure 11C**, $r = 0.02$, $p\text{-value} = 0.8288$).

The tumour DNAm age does not differ significantly according to the smoker status (**Figure 11E**, $p\text{-value} = 0.1908$).

Tumour age acceleration

The tumour age acceleration does not correlate with the number of years smoking (**Figure 11B**, $r = 0.12$, $p\text{-value} = 0.1123$).

It does not correlate with the number of cigarette packs smoked per year either (**Figure 11D**, $r = -0.01$, $p\text{-value} = 0.8593$).

However, the tumour age acceleration is significantly different in the smoker population vs. the non-smoker population (**Figure 11F**, p-value = 0.02594). The non-smoker population has a mean tumour age acceleration of 0.61 while the smoker population's mean is 0.54. Although not easily interpreted, this result is interesting. It would indicate that smoking somewhat increases the difference between the chronological age of the patient and the DNAm age of his/her tumour, which appears on average around twice younger than the patient him/herself.

Conclusion

The tumour's DNAm age weakly, but significantly, correlates with the number of years the patient has been smoking: if the patient smoked for a longer amount of time, his/her tumour tends to appear "older". However, it does not seem affected by the amount of tobacco smoked during this time nor by the smoker status as a binary variable. Tumour age acceleration does not seem to be affected nor by the number of years smoking nor by the number of packs smoked per year, but it does seem to be affected by the smoker status of the patient: smokers tend to have a lower tumour DNAm age/chronological age ratio. These results are somewhat paradoxical: the tumour seems to appear older in the first case, but seems younger (when normalised to account for the patient's age) in the second.

Tobacco is, with alcohol consumption, the most important risk factor for HNSC cancer⁹. Although results are hard to interpret, they show this might be reflected in the apparent "age" of the tumour to some extent. Nonetheless, the interpretation for the DNAm age vs. years smoking correlation could also be more trivial: if the patient has been smoking for a higher number of years, he would tend to be older, and thus his/her tissues, including the tumour, would tend to appear older in terms of DNAm. The chronological age of the patient and the amount of years he/she has been smoking highly correlate (figure not shown, $r = 0.75$, p-value ≈ 0). However, the amount of years smoking does not seem to affect the age of the normal tissues (figure not shown, $r = 0.19$, p-value = 0.3777). Sample size is low though ($N = 24$ patients), so this does not necessarily confirm, nor refute, the conclusion that the amount of years smoking directly "ages" the tumour. Overall, this explanation may indicate that the second result, i.e. that smokers tend to have younger tumours compared to their chronological age (lower age acceleration), should perhaps be more trusted than the first. Indeed, it is unlikely to be biased by chronological age thanks to the definition of age acceleration.

The precise anatomical location where primary tumour is located is not mentioned in this file. If it was, it might have been possible to reach more meaningful conclusions by sorting the patients based on this information. Indeed, tobacco particularly increases the risk of HNSC tumours located in the oral cavity, oropharynx, hypopharynx, and larynx, but it does not, for instance, increase the risk for salivary gland cancer⁹.

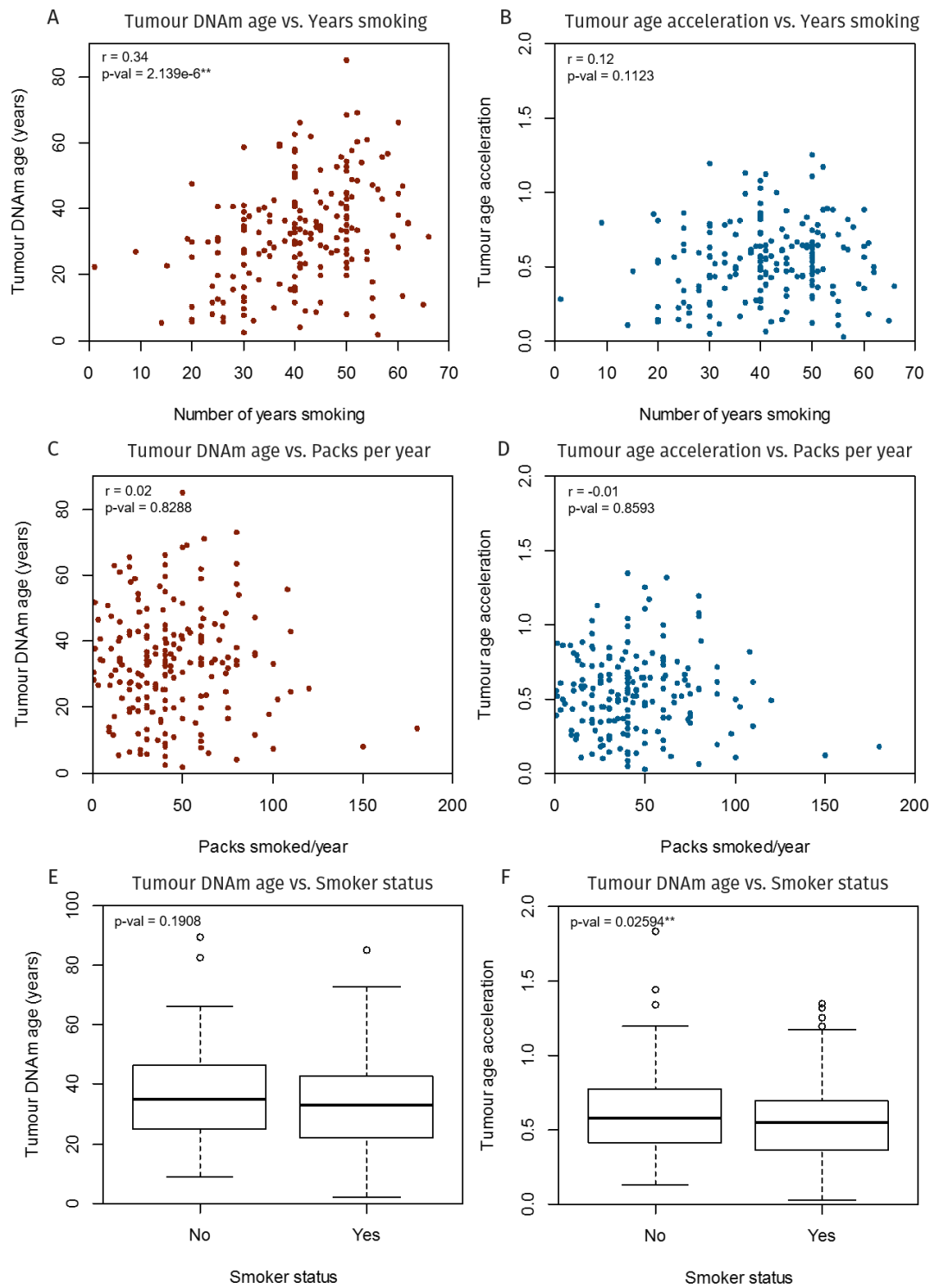


Figure 11 | Tumour age vs. smoking behaviour. **(A)** Tumour DNAm age vs. number of years smoking (from the smoking onset to the initial pathologic diagnosis) scatter plot. **(B)** Tumour age acceleration vs. number of years smoking scatter plot. **(C)** Tumour DNAm age vs. the average number of packs smoked per year scatter plot. **(D)** Tumour age acceleration vs. the average number of packs smoked per year scatter plot. One patient who smoked an average of 300 packs per year was omitted from plots C and D for graphical reasons. **(E)** Tumour DNAm age vs. the smoker/non-smoker patient status boxplot. **(F)** Tumour age acceleration vs. the smoker/non-smoker patient status boxplot. Spearman's correlation coefficient and its p-value is included in each plot (A–D). The p-value computed by Welch's t-test is included in each plot (E, F).

Other clinical variables

Several other clinical variables from the file *All_CDEs.txt* were tested for relationships with the tumour's DNAm age or age acceleration. As these variables are not the primary goal of this section, there are more succinctly presented in **Table 1**.

Table 1 | Additional clinical variables

Variable	Type & N (total = 348)	vs. tumour DNAm age	vs. tumour age acceleration
<i>Disease after curative treatment</i>	Binary (15 yes, 110 no, 223 NA)	p-val = 0.3813	p-val = 0.07732
<i>Follow-up treatment success</i>	Categorical (5 outcomes, 139 NA)	p-val = 0.304	p-val = 0.204
<i>Alcohol history documented</i>	Binary (241 yes, 100 no, 7 NA)	p-val = 0.3668	p-val = 0.9798
<i>Frequency of alcohol consumption</i>	Continuous (196 NA)	r = -0.09 p-val = 0.2641	r = -0.06 p-val = 0.4553
<i>Amount of alcohol consumption per day</i>	Continuous (199 NA)	r = -0.01 p-val = 0.8753	r = 0.04 p = 0.5973
<i>HPV status</i>	Binary (71 positive, 276 negative, 1 NA)	p-val = 0.1296	p-val = 0.5024
<i>HPV type</i>	Categorical (5 types, 278 NA)	p-val = 0.213	p-val = 0.23

Conclusions

Nor the tumour's DNAm age nor its age acceleration seem to significantly affect the outcome of the treatment.

Nor alcohol consumption nor infection by the human-papillomavirus (HPV) seem to significantly affect the tumour DNAm age or age acceleration, although these two variables are strong risk factors for HNSC cancer, especially infection with the HPV type 16⁹.

Q2.2: Survival analysis

This section will explore survival data in the aim of answering whether patient's survival is affected by the tumour DNAm age.

Material and methods

Survival data are included in the main clinical file used from Q1 (*HNSC.clin.merged.picked.txt*, from *Clinical_Pick_Tier1*³).

Survival analysis makes use of the R package *survival*¹¹. It uses three clinical variables: the *vital status* (1 if the patient died, 0 if he/she is alive at the time of the last follow-up) and *days to death/days to last follow-up*, which represent the number of days between diagnosis and date of death or last follow-up. These data do not overlap, i.e. any patient who has a number in *days to last follow-up* was alive at the time of last follow-up, and thus does not have a number in *days to death*, and vice-versa. These two columns were merged into one single column named *days*. All patients have a number of days in this column (348 number of days, no missing data).

From this column, we simply added a + sign to the number of days of patients still alive at the time of the last follow-up (i.e. patients whose vital status is 0). This is done with the command:

```
Clinical$survival <- with(Clinical, Surv(days, vital_status == 1))
```

where *Clinical* is our main data frame containing data from the two clinical files and DNAm age file. *Surv* is a function from the *survival* package¹¹. The + sign (e.g. 230+) indicates that the number of days for this specific patient is right-censored. Right-censoring means that we stopped keeping track of this patient: he/she was still alive at the time of last follow-up, but we did not follow what happened to him/her after this. If the number of days is not right-censored (i.e. there is no +), then it represents the actual survival time, i.e. the number of days the patient survived after diagnosis.

These data are plotted as a Kaplan-Meier plot using:

```
plot (survfit (survival ~ 1, data = Clinical), mark.time = TRUE, mark = 'line')
```

where *survfit* is a function of the *survival* package. *mark.time* and *mark* allow us to add lines across the curve to indicate the right-censored data.

Kaplan-Meier curves are a specific type of plots designed to display survival data. It plots the proportion of patients alive as a function of time. For instance, if the curve is at 0.4 on the Y axis and at 2000 days on the X axis, it should be read as: "at 2000 days after diagnosis, 40% of the initial population of patients are still alive".

More interestingly, Kaplan-Meier curves can be used to visualise if survival differs between two populations, for instance gender. This can be obtained by replacing the 1 in the command here above. For instance, for gender:

```
plot (survfit (survival ~ gender, data = Clinical), ...)
```

In practice, graphical parameters (...) and a legend are added to allow the reader to differentiate the two curves (e.g. **Figure 12**).

Survival data can also be statistically tested. If we wish to test them against a categorical variable (e.g. gender or pathologic stage), i.e. we want to compare if survival significantly differs in k different populations, we typically use a log-rank test, also called a Cox regression. We can also use the Cox regression to test the survival data against a continuous variable (e.g. weight or height). However, in the latter case, the data cannot be easily represented with a Kaplan-Meier plot, as it would involve creating arbitrary categories. Finally, the combined effect of two or more variables can also be tested at once, which is called a Cox multiple regression.

In addition to the p-value, the Cox regression analysis provides us with the ratio of the hazards (HR), called *exp(coef)* in R. This allows to interpret the magnitude of the effect of the variable on survival. For a categorical binary variable (e.g. gender), it compares the “risk of an endpoint”, i.e. death, between the two population (example hereafter). For a continuous variable, it describes by how much the risk is increased or decreased when the variable is increased by one unit. It is called the hazard ratio because it is formally defined as the ratio between the hazards of two patients differing only by one unit of the variable (for instance, male vs. female in case of a categorical variable; or 79 vs. 80 kg in case of a continuous variable). For example (not actual data), if the hazard ratio is 1.05 when survival is tested against weight, it means that the risk of dying is multiplied by 1.05 each time weight increases by one kilogram. The HR’s 95% confidence interval (CI) is also provided. Intuitively, it allows to estimate the range in which we can be almost (95%) sure the “real” HR is, i.e. if sample size was infinite.

For example, in our data, survival is significantly different in the female vs. male population (**Figure 12A**, p-value = 0.034): men survive longer in average to HNSC cancer than women. The HR for male is 0.6903. Stated differently, female patients are in average $1/0.6903 = 1.45$ times more likely to die at any given time.

Another example is radiation therapy. Patients who receive radiation therapy survive significantly better than those who do not (**Figure 12B**, p-value = 0.00345). The HR for patients who received radiation therapy is 0.5882. Intuitively, it means that a patient who did not receive radiation therapy is $1/0.5882 = 1.7$ times more likely to die at each time unit.

On a side note, the Cox regression makes some underlying assumptions about the data. The most important is the proportional hazard assumption. It can graphically be assessed by looking at the two curves: if they are approximately parallel, the proportional hazard assumption can safely be made. In our plots (**Figure 12**), we can see that this assumption seems to hold until ~ 2200 days in the gender plot and until ~ 2800 days in the radiation therapy plot. As in both cases, the number of patients left and the number of events (deaths) occurring after this point are low,

these data points will contribute less to the estimates of the Cox regression. Overall, the conclusions of the Cox regressions are likely to be correct.

Tumour DNAm age

The tumour DNAm age does not seem to affect survival (p-value = 0.538, HR = 0.9968, CI = 0.9866–1.007).

Tumour age acceleration

The tumour age acceleration does not seem to affect survival (p-value = 0.125, HR = 0.6276, CI = 0.3463–1.138).

On a side note, if definition 1 of the tumour age acceleration is used, it seems to significantly worsen survival (p-value = 0.04078, HR = 2.875, CI = 1.041–7.942). However, the 95% CI is huge and almost includes 1.0 (i.e. no effect on survival), these results should thus probably not be relied upon.

Tumour DNAm age + age acceleration

As mentioned, Cox regression can also be used with a combination of variables. In this case, it measures the combined effect of the variables on survival, in our case the tumour's DNAm age and the tumour's age acceleration. When combined with the tumour's DNAm age, the age acceleration seems to significantly improve survival (p-value = 0.0276, HR = 0.17344, CI = 0.03649–0.8245). The HR would indicate a protective effect of the age acceleration on survival: the higher the age acceleration, the higher the survival probability. However, this conclusion should be interpreted with some caution: first, the 95% CI is quite large, although it does not include 1.0; second, the age acceleration also includes in its computation the tumour's absolute DNAm age, and using twice this variable might be erroneous in some sense; third, if the effect of the tumour's age, in absolute or relative (acceleration) terms, was particularly strong, it should probably have been detected using the DNAm age or the age acceleration alone.

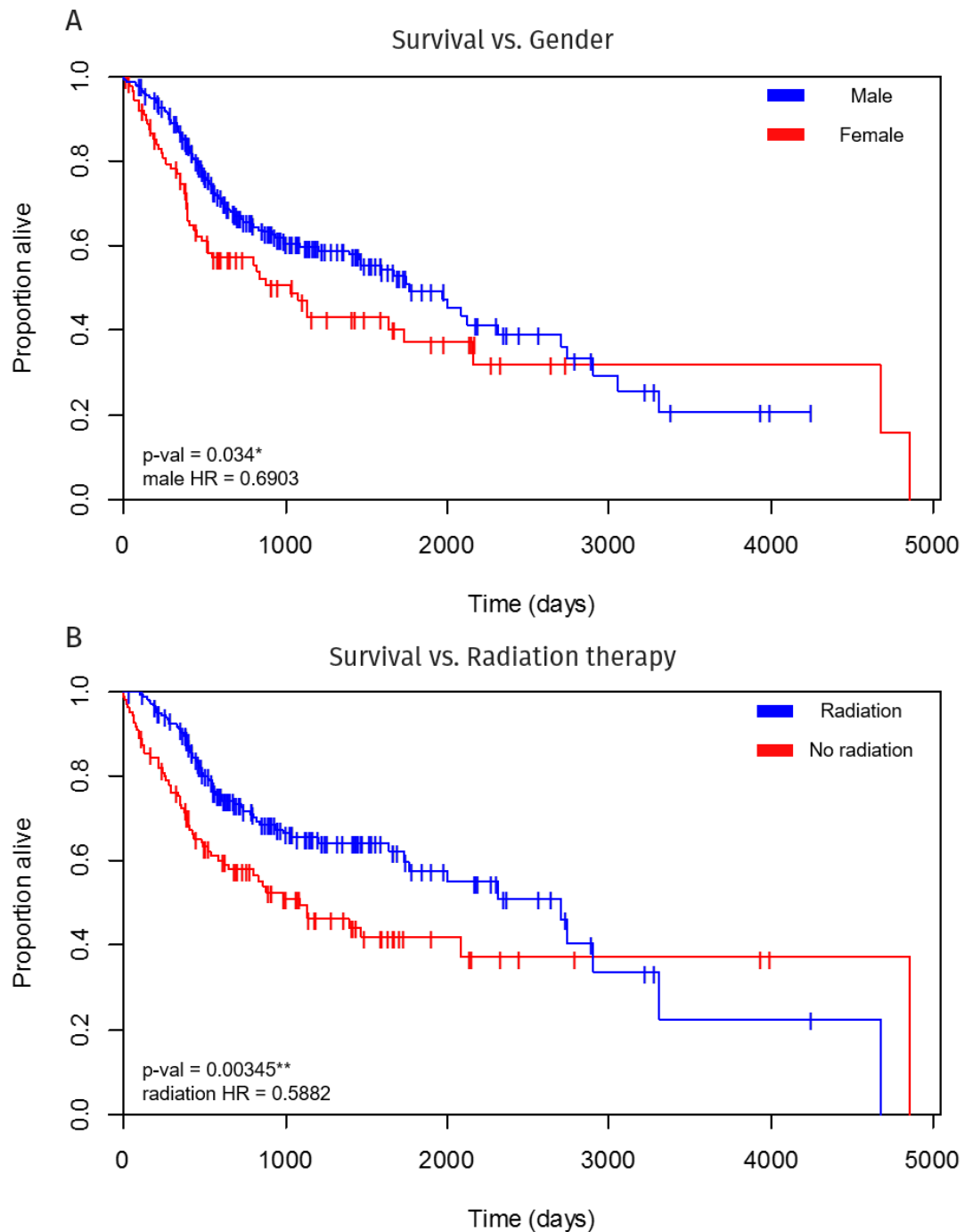


Figure 12 | Examples of Kaplan-Meier curves. **(A)** Survival vs. gender (male in blue vs. female in red). **(B)** Survival vs. whether the patient received radiation therapy or not (yes in blue vs. no in red). Vertical lines represent right-censored data. The p-value computed by Cox regression is included in each plot. N = 348 for both.

Q3: Gene expression

This question will explore RNA sequencing (RNA-Seq) data in the aim of exploring how the tumour DNAm age affects the cancer cells' gene expression.

The first section will explore gene expression data where genes are classified in gene sets, i.e. which, if any, sets of genes are up- or down-regulated when the tumour's age increases? The second section will investigate gene expression at the gene scale, i.e. which, if any, single genes are up- or down-regulated when the tumour's age increases?

Material and methods

Gene sets

The first section about gene sets uses the gene set enrichment analysis (GSEA) software¹² from the Broad Institute and UC San Diego, in the Java desktop application form (v2.2.4)¹. GSEA needs three types of data as input: gene expression data; phenotype data; and a database of the gene sets.

In our case, the gene expression data are the normalised RNA-Seq counts. The file GSEA used was prepared from the normalised RNA-Seq counts downloaded from FireHose (*HNSC.rnaseqv2_illuminaiseq_rnaseqv2_unc_edu_Level_3_RSEM_genes_normalized_data.data.txt*, from *illuminaiseq_rnaseqv2-RSEM_genes_normalized* in the *mRNA-Seq* section). Data from this file needed some filtering. Briefly, there were RNA-Seq data from 2 metastases, 44 healthy tissues, and 520 primary tumours. First, data from the metastases and the healthy tissues were removed: they should not be analysed together with the primary tumour samples, and we are mainly interested in the cancer cells' gene expression. Second, we are exploring the possible effects of the tumour DNAm age on gene expression. It is thus necessary that we have DNAm age data for the patients we are going to analyse. From the 520 patients whose tumour was sequenced, only 342 also had a tumour DNAm age, i.e. were also present in the clinical data frame. We thus removed from the RNA-Seq the patients who did not have a DNAm age (178 patients), and from the clinical file the patients who did not have their tumour sequenced (6 patients). Both files were ordered alphabetically according to the patient's code so we could be assured that a column in the RNA-Seq file and the row of same index in the clinical file correspond to the same patient. This matching is crucial to our analysis and was thoroughly checked. In the RNA-Seq file, each gene has two names separated by a pipe: first its ENTREZ ID, then its GENE SYMBOL (e.g. A1BG|1). GSEA does not recognise this notation and only uses one or the other. Using a regular expression pattern, we kept only the ENTREZ ID part. The first 29 genes (rows) had unknown gene ENTREZ IDs (denoted as "?" symbol); these rows were removed. There was also one duplicate in the genes, whose row was also removed. The final RNA-Seq data frame contains 20501 rows, each representing a gene; and 342 columns, each representing a patient (his/her primary tumour). Finally, the RNA-Seq file and a file containing the matched tumour DNAm ages were exported from R as text files to be edited in Excel, which is advised by

the GSEA's authors to prepare the precise files formats¹³. Without entering into details, this editing mainly involves adding a couple of lines at the beginning of the table mentioning the number of samples (342) and the number of genes (20501), and a column which can be used to include a description of each gene. This column was filled by dummy NA values in our case. So it can be recognised by GSEA, the format of the final file must be changed from *.txt* to *.gct*. Similarly, a couple of lines also needed to be added with Excel to the DNAm ages file. The only purposes of these lines are to mention that the variable is continuous (with *#numeric*) and its name (*#tumour DNAm age*). The final file format is *.cls*.

These two files were then loaded in GSEA. The gene set database was directly selected from GSEA. It was *c2.cp.v6.0.symbols.gmt*¹². Selected parameters are included in **Figure 13A**. Default parameters were used, except for the *Collapse dataset to gene symbols*, which was set to *false*, and the *Metric for ranking genes*. Our phenotype (DNAm age or age acceleration) being continuous, it needs a metric which functions with continuous variables. We chose Pearson's correlation.

Succinctly, GSEA will first rank our list of 20501 genes based on this metric: the better the gene's expression correlates with the tumour DNAm age, the higher it is ranked in this list. The top gene in the list is the gene whose expression positively correlates the highest with the phenotype, here tumour DNAm age. The last gene in the list is the gene whose expression negatively correlates the highest with tumour DNAm age, and somewhere in the middle lie genes whose correlations are around 0, i.e. their expression is not affected at all by the phenotype. It will then walk down the ranked list, keeping a running-sum statistic. At each gene, GSEA increases the running-sum statistic of the set(s) which include(s) this gene, and decreases the running-sum statistic of the sets that do not. The final enrichment score (ES) is the maximum score the running-sum statistic reached during the walk down. The enrichment plot in the GSEA results summarises this process (e.g. **Figure 13B**). The upper part is the enrichment profile: it represents how the running-sum statistic evolved throughout the walking down process. The middle part represents the hits, i.e. the genes of the set that were matched during the walking down, and the colour scale and the lower part represent the ranked list of genes GSEA walked down¹⁴.

The number of permutations represents the number of times GSEA will randomly shuffle the phenotypes, i.e. randomly assign the DNAm ages of tumours to other tumours. This is used to estimate the false discovery rate (FDR). A commonly used threshold for the FDR is 25%: if a gene set is significantly (p-value < 0.05) enriched in the initial data, but was also significantly enriched in more than 250 of the 1000 shuffles, it has > ¼ chance to be a false positive. It should thus not be trusted, even though its nominal p-value is low.

Single genes

The analysis of single genes expression was performed using the *samr* R package¹⁵.

SAM (significance analysis of microarrays) is a package that allows to correlate the expression of a large number of genes with an outcome variable, in our case the tumour DNAm age or age

acceleration. As its name suggests, it was originally designed for microarray data, but it can now be used for RNA-Seq data using a command like:

```
samfit <- SAMseq (rnaseq_tumor, tumor_age, resp.type = 'Quantitative')
```

where *rnaseq_tumor* is a matrix containing the normalised number of counts. It is basically the same than the data frame used for the gene sets: each row (20501 rows) represents a gene and each column (342 columns) a patient (his/her primary tumour). For some reason, SAM does not seem to manage non-integer values, and thus the normalised counts were rounded to the closest integer. *tumor_age* is a simple vector containing the DNAm age or the age acceleration of each tumour. When preparing the files for GSEA, patients were alphabetically ordered in both the main clinical file and the RNA-Seq data file before the tumours' ages were extracted. Therefore, we can be sure a specific tumour has the same index in the RNA-Seq file and the tumour ages vector.

Gene sets

Tumour DNAm age

Here, we are assessing whether the tumour DNAm age is significantly correlated with an up- or down-regulation of specific set(s) of genes.

Out of 1070 gene sets in the database, 454 are up-regulated, i.e. their expression correlates with DNAm age. Among them, 41 are significantly enriched with a p-value < 0.05. Nevertheless, none had a FDR below 25%, so these results should not be trusted as they are likely to be false positives.

For information, the two up-regulated pathways with the lowest nominal p-values are the set of genes involved in asthma (*KEGG_ASTHMA*, 28 genes, **Figure 13B**) and the set of genes involved in autoimmune thyroid disease (*KEGG_AUTOIMMUNE_THYROID_DISEASE*, 50 genes). Although not significant when the p-value is corrected to account for the multiple hypotheses (family-wise error rate (FWER) = 0.539 and 0.514, respectively), it may still be linked somehow to HNSC cancer as asthma may protect against oropharyngeal cancer (one type of HNSC cancer)¹⁶, and autoimmune thyroid disease is a potential side effect of radiation therapy¹⁷.

Oppositely, out of 1070 gene sets in the database, 616 are significantly down-regulated with the profile, which means their expression is negatively correlated with tumour DNAm age. Stated differently, if tumour DNAm age is high, their expression tends to be low. Among them, 42 have a nominal p-value below 0.05, and more importantly, 3 have FDRs below 25%: the set of genes involve in the pentose phosphate pathway (27 genes), the gluconeogenesis set (31 genes) and the broader glucose metabolism set (64 genes). The first observation is that these three sets are involved in glucose metabolism. The set of genes involve in the pentose phosphate pathway is the most interesting: it is the set which has the lowest nominal p-value (≈ 0), the lowest FDR (0.170352), and the lowest FWER (0.084). Although not significant, the FWER is not far from the threshold of 0.05, and is considerably lower than just the second-best set (FWER = 0.143).

Although this is certainly true for other types of cancer, this pathway, together with glucose metabolism altogether, are mentioned in literature alongside HNSC cancer^{18 19}. The HNSC cells are typically highly dependent on glucose, and quickly die in glucose-free medium. What is surprising here is that expression of genes involved in the glucose metabolism seems to be negatively correlated with the tumour DNAm age. This may represent an interesting difference in metabolism between “young” and “old” HNSC tumours. Except if this decrease in glucose metabolism was compensated by other pathways, this could also indicate a decreased “vigour” of the tumour when it appears older. An interesting *in vitro* experiment here could be to test how these cancer cells which appear “old” react in a glucose-free medium vs. glucose-enriched medium.

Also for information, the gene whose expression correlated the most positively with the tumour DNAm age is PTX4 (pentraxin-4, $r = 0.34$). Pentraxins are an evolutionary conserved family of proteins whose members are mainly involved in acute immune function. The gene whose expression correlated the most negatively with the tumour DNAm age is MAGEF1 (melanoma-associated antigen F1, $r = -0.32$). Expression of group 1 MAGEs is restricted to tumours and germ cells²⁰.

Tumour age acceleration

Among 1070 gene sets, the expression of 471 positively correlated with the tumour age acceleration. 14 of them have a p-value below 0.05 but none have a FDR below 25%. For information, the gene set with the lowest nominal p-value is the set of genes involve in the calcium homeostasis in platelet (p-value = 0.005769). It is also the gene set with the lowest FWER (0.762), but has a FDR of 100%. It is thus almost certain to be a false positive.

Among 1070 gene sets, the expression of 599 negatively correlated with the tumour’s age acceleration. 27 of them have a p-value below 0.05, and one has a FDR below 25%. Interestingly, the four pathways with the lowest FDR are exactly the same than in the case of the DNAm age: the pentose phosphate pathway, the glucose metabolism, the gluconeogenesis and the glycolysis. The pentose phosphate pathway has the lowest FDR (0.225513), the second lowest nominal p-value (0.004211) and the third lowest FWER (0.228).

Conclusion

The safest conclusion here would be that no gene set were significantly up- or downregulated as a function of tumour DNAm age or age acceleration in terms of FWER (no gene set had a FWER < 0.05). However, the apparent downregulation of glucose metabolism in older tumours is very interesting as these tumours are paradoxically known to rely critically on glucose^{18 19}. Some of these gene sets have FDR below 25%. Additionally, the fact that these four gene sets (pentose phosphate pathway, glucose metabolism, gluconeogenesis and glycolysis) are the top four most negatively correlated pathway both as a function of DNAm age and age acceleration, and in terms of nominal p-value, FDR and FWER in both cases, is very likely to be meaningful.

A

Required fields

Expression dataset: maseq_tumor [20501x342 (ann: 20501,342,chip na)]

Gene sets database: roadinstitute.org/pub/gsea/gene_sets_final/c2.cp.v6.0.symbols.gmt

Number of permutations: 1000

Phenotype labels: c:\pbox\FunctionalGenomics project\tumor_age.cls#tumor DNAm age

Collapse dataset to gene symbols: false

Permutation type: phenotype

Chip platform(s):

Basic fields Hide

Analysis name: maseq_tumour

Enrichment statistic: weighted

Metric for ranking genes: Pearson

Gene list sorting mode: real

Gene list ordering mode: descending

Max size: exclude larger sets: 500

Min size: exclude smaller sets: 15

Save results in this folder: C:\Users\François\gsea_home\output\juin13

Advanced fields Show

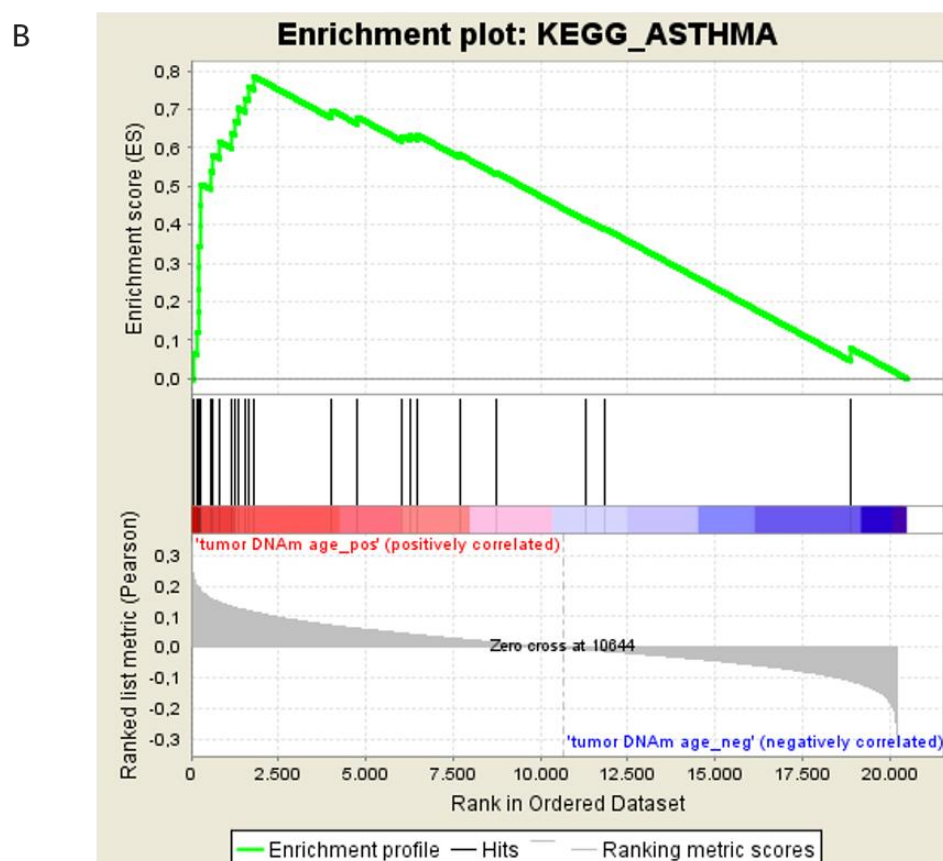


Figure 13 | Using GSEA to correlate the expression of gene sets with the tumour age. **(A)** Parameters used when running GSEA. Parameters that were modified from default ones are squared in red. No parameter was changed from default in the *Advanced fields* section. **(B)** Example of GSEA enrichment plot. The upper part is the enrichment profile, while the lower part represents the ranked gene list.

Single genes

We are assessing here which genes are significantly positively or negatively correlated to the tumour DNAm age. 20501 genes were studied so the complete results cannot be included in this report. We will only look at the top-three up- and down-regulated genes as a function of tumour DNAm age or age acceleration.

Tumour DNAm age

Table 2 summarises the top-three up-regulated genes as a function of tumour DNAm age. **Table 3** summaries the top-three down-regulated genes as a function of tumour DNAm age. **Figure 14A** displays the differentially expressed genes relative to tumour DNAm age.

Table 2 | Up-regulated genes vs. tumour DNAm age.

Top-3 Up-regulated genes			
Entrez ID	Gene symbol	t-value	FDR
4159	MC3R	-0.388	0
17011	Ltrm1(<i>Mus musculus?</i>)	-0.332	0
3159	HMGA1	-0.309	0

Table 3 | Down-regulated genes vs. tumour DNAm age.

Top-3 Down-regulated genes			
Entrez ID	Gene symbol	t-value	FDR
3719	JBS (?)	0.374	0
10480	EIF3M	0.373	0
6357	CCL13	0.358	0

Tumour age acceleration

Table 4 summarises the top-three up-regulated gene as a function of tumour age acceleration. **Table 5** summaries the top-three down-regulated gene as a function of tumour age acceleration. **Figure 14B** displays the differentially expressed genes relative to the tumour age acceleration.

Table 4 | Up-regulated genes vs. tumour age acceleration.

Top-3 Up-regulated genes			
Entrez ID	Gene symbol	t-value	FDR
13350	Dgat1 (<i>Mus musculus?</i>)	-0.388	0
4159	MC3R	-0.332	0
3159	HMGA1	-0.309	0

Table 5 | Down-regulated genes vs. tumour age acceleration.

Top-3 Down-regulated genes			
Entrez ID	Gene symbol	t-value	FDR
3719	JBS (?)	-0.388	0
10480	Ltrm1(<i>Mus musculus?</i>)	-0.332	0
15845	Iapls3-37 (<i>Mus musculus?</i>)	-0.309	0

Conclusion

Except from the fact that there are some common genes in the most up- and down-regulated genes as a function of tumour DNAm age or age acceleration, it is difficult to draw meaningful conclusions from these lists of genes. SAM was initially designed for analysis of microarray experiments results, where the genes probes are typically manually selected. In this case, we know which genes we are specifically interested in and it is easy to look them up in the list. The fact that we compare expression as a function of a continuous variable complexifies also the interpretation.

Overall, these results mainly indicate that SAMseq might not be the ideal tool in our setting. As we have seen, analysis of enriched pathways can be more much productive when that many genes (20501) are analysed in in that many conditions (342 conditions, i.e. patients).

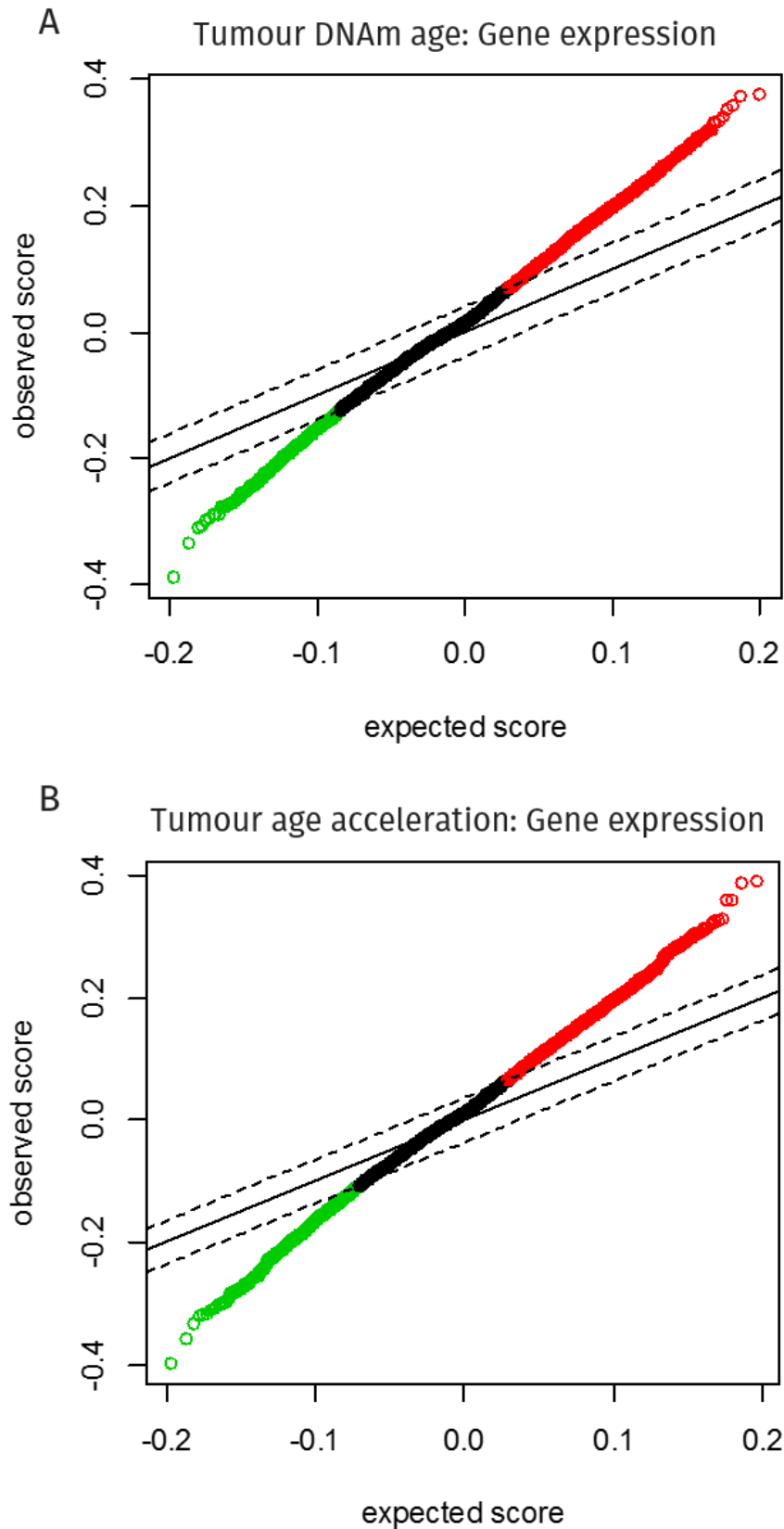


Figure 14 | SAMseq results for differentially expressed genes. Each gene is represented by a dot. It is in red if it is up-regulated as a function of tumour age, in green if it is down-regulated as a function of tumour age, and in black if its expression does not change as a function of tumour age. **(A)** Tumour age is defined as absolute DNAm age. **(B)** Tumour age is defined as relative age acceleration. N = 20501 genes in both plots.

Bonus question 2: Copy number variations

This section is more experimental and should be taken with more caution.

Material and methods

The analysis of the copy number variations uses as input the file *genome_wide_snp_6-segmented_scna_minus_germline_cnv_hg19* from FireHose³ (section *SNP6 CopyNum*).

For each sample (patient), an SNP array has been run on the patient's tumour and on a sample of matched normal tissue (i.e. from the same anatomical site). *Minus germline* indicates that from the raw number of copies from the tumour sample has been subtracted the number of copies of the same sequence in the normal tissue. The goal is to have meaningful data to analyse copy number variations in tumours: a positive number indicates duplication events, i.e. the sequence is more repeated in the tumour than it was in the normal tissue; a negative number indicates deletion events, i.e. the sequence is less repeated in the tumour than it was in normal tissue. Furthermore, the copy number data included in the file are in the $\text{Log}_2\left(\frac{\text{Copy number}}{2}\right)$ form, which we will call copy number variation (CNV). The goal of this transformation is to normalise the data (Log_2) and account for diploidy (*divided by 2*).

The original file contained 110289 rows. Among them, 30927 were data from normal blood samples (tissue code *10*), 74970 from primary tumour samples (tissue code *01*), 4011 from normal tissue samples (tissue code *11*) and 381 from metastases (tissue code *06*). Only data from the primary tumours were kept (74970 rows).

Furthermore, the original file contained data about 526 individual patients. For **Figure 17**, we needed to keep only the patients that were also present in the clinical data frame (see previous questions), i.e. had a tumour DNAm age. 345 of the 526 were: their data were kept, while the data from the other patients were removed. Then, the data for each unique patient had to be separated from the original file to be averaged. Very briefly, this was done by first splitting the original file into the data for each individual patient. The results were stored in a list of matrices. The CNV data was then averaged in each single matrix (one matrix per patient), and stored in a vector. Both datasets were sorted in alphabetical order of the patient code so we could be sure to match the average CNV with the correct tumour.

Single patient example

A first example is plotted to better understand how the data works (**Figure 15**). It displays the CNV data of the first patient of the file (first 280 rows) as a function of the genomic positions. Each data point represents a DNA sequence, its X coordinate is its start position on the reference human genome (*hg19* version), and its Y coordinate is the CNV. Sequences whose CNV is ~ 0 (red

line) did not undergone duplication(s) or deletion(s). Sequences whose CNV is above the red line are duplicated in the tumour compared to normal tissue, and sequences whose CNV is below the red line are deleted in the tumour compared to normal tissue.

It should however be noted that genomic positions are not unique. For example, chromosome 1 and chromosome 23 will both start at position 1. Thus, we should keep in mind that this plot (**Figure 15**) overlays all 23 chromosomes on the same graph.

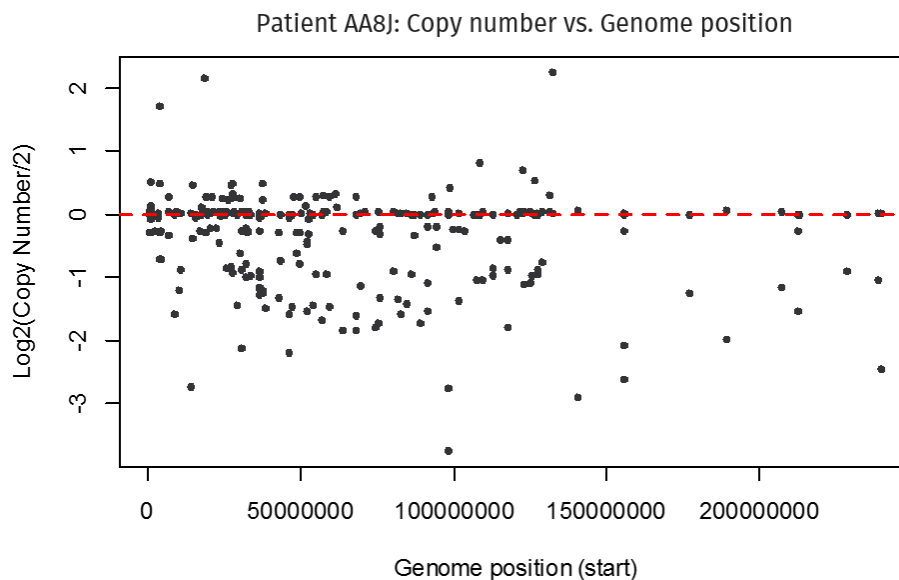


Figure 15 | Copy number variations (CNV) vs. genome start positions for patient AA8J (N = 280 sequences).

All patients

We should now try to plot these data for all patients at once. A similar plot should probably be used. If some specific positions are consistently duplicated/deleted in the tumour compared to normal tissue, it should appear as a high point-density zone in the plot.

A solution could be to simply overlay the data for all the patients on the same plot (**Figure 16**).

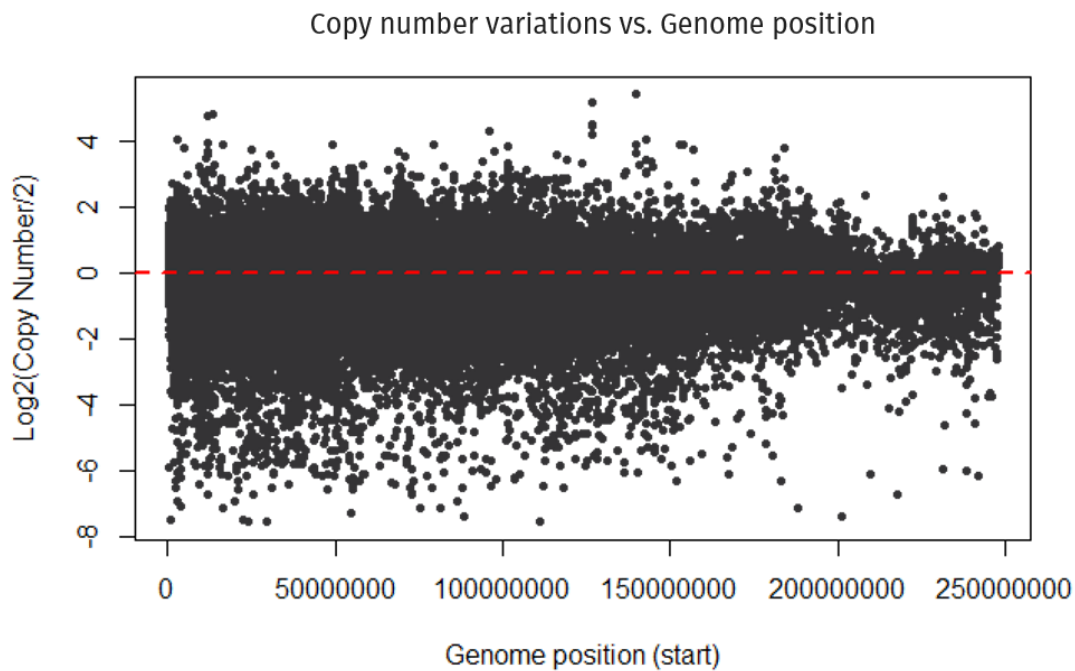


Figure 16 | Copy number variations vs. genome positions (start position of each sequence) for all patients' primary tumours (N = 74970 sequences, 526 patients).

Copy number variation vs. tumour age

While this allow us to get more insight into what the data include, it does not answer the question as whether the copy number variation correlates with the tumour age.

The problem here is to decide what is the most correct way to simplify and display this huge amount of data.

A first rough solution here could be to take the average CNV for each patient's primary tumour. Based on **Figure 15** and **Figure 16**, there seem to be more sequences deleted than duplicated during tumour development, so the means are likely to not be around 0 and could be meaningful when plotted against tumour age.

This is done in **Figure 17**. Each data point (N = 345) thus corresponds to one patient. Its X coordinate is the age of his/her tumour and the Y coordinate is the average CNV across all sequences from the tumour cells.

There is an extremely weakly negative ($r = -0.15$), but highly significant ($p\text{-value} = 0.005023$), correlation between the tumour DNAm age and its average CNV (**Figure 17A**).

There is also an extremely weakly negative ($r = -0.17$), but highly significant ($p\text{-value} = 0.00113$), correlation between the tumour age acceleration and its average CNV (**Figure 17B**).

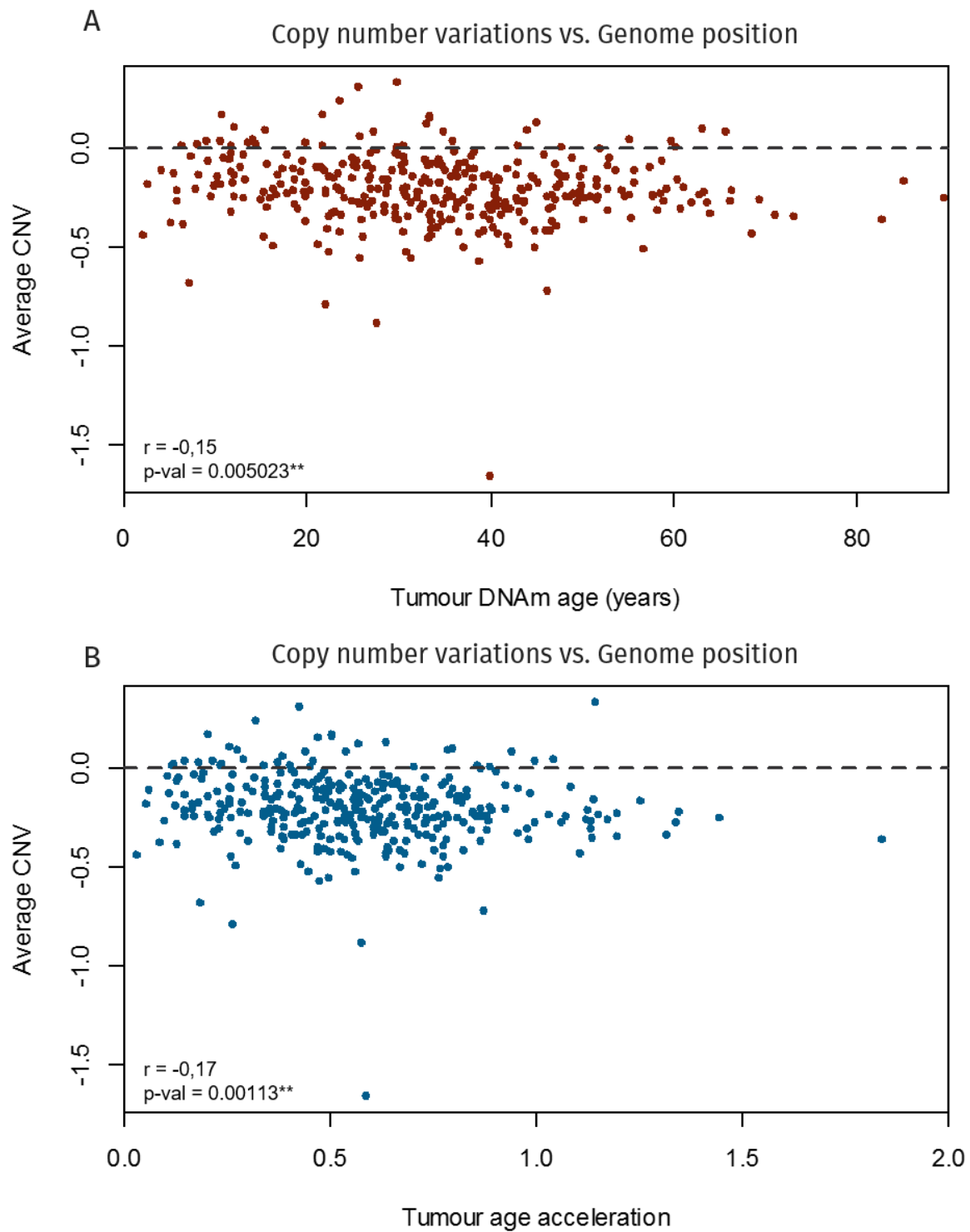


Figure 17 | Average CNV per tumour vs. tumour age scatter plot. **(A)** vs. tumour DNAm age. **(B)** vs. tumour age acceleration. Spearman's correlation coefficient and its p-value is included in each plot. N = 345 tumours/patients for both plots.

Conclusion

At first approximation, tumour age may be slightly correlated with deletion of sequences, i.e. a decrease in CNV. While these correlations are not strong, these plots allow at least to confirm why we anticipated previously: the CNV of tumour cells is negative in average. This means that during tumour development, sequences are deleted.

Finally, we plotted here the average CNV per tumour. Thus, there might be very interesting and significant results lying in specific sequences, that could be specifically duplicated or deleted as a function of tumour DNAm age/age acceleration. The challenge remains to spot these sequences. Once this is done, a tool like the UCSC Genome Browser²¹ could be used to see if the duplicated/deleted sequences which correlate the most (positively or negatively) with tumour age represent specific genes. Indeed, the CNV data include the chromosome number and the start and end of each sequence on the hg19, which is all you need to perform a query in the Genome Browser.

Bibliography

1. BroadInstitute. Gene Set Enrichment Analysis (GSEA). Available at: <http://software.broadinstitute.org/gsea/index.jsp>. (Accessed: 13th June 2017)
2. R Core Team. R: A Language and Environment for Statistical Computing. (2017).
3. FireHose - Broad Institute. HNSC Archives. Available at: http://firebrowse.org/?cohort=HNSC&download_dialog=true. (Accessed: 11th June 2017)
4. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, 1–19 (2013).
5. NIH - The Cancer Genome Atlas. About TCGA Data. Available at: <https://tcga-data.nci.nih.gov/docs/publications/tcga/about.html>. (Accessed: 14th June 2017)
6. University of Virginia. Understanding Q-Q Plots. Available at: <http://data.library.virginia.edu/understanding-q-q-plots/>. (Accessed: 14th June 2017)
7. UCLA - Institute for Digital Research and Education. What is the difference between categorical, ordinal and interval variables? Available at: <https://stats.idre.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/>. (Accessed: 14th June 2017)
8. Stevenson M., M. Head and Neck Cancer Staging. *Medscape* (2016). Available at: <http://emedicine.medscape.com/article/2007181-overview>. (Accessed: 12th June 2017)
9. National Cancer Institute. Head and Neck Cancers. (2017). Available at: <https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet>. (Accessed: 12th June 2017)
10. Yu, M. & Yuan, J. Epidemiology of nasopharyngeal carcinoma. *Semin. Cancer Biol.* **12**, 421–429 (2002).
11. Therneau, T. M. A Package for Survival Analysis in S. (2015).
12. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**, 15545–15550 (2005).
13. Broad Institute. Data formats. Available at: http://software.broadinstitute.org/cancer/software/gsea/wiki/index.php/Data_formats. (Accessed: 13th June 2017)
14. Broad Institute. GSEA User Guide. Available at: <http://software.broadinstitute.org/gsea/doc/GSEAUUserGuideFrame.html>. (Accessed: 15th June 2017)
15. Tibshirani, R., Chu, G., Narasimhan, B. & Li, J. samr: SAM: Significance Analysis of Microarrays. (2011).
16. Michaud, D. S. *et al.* Allergies and risk of head and neck cancer. *Cancer Causes Control* **23**, 1–10 (2012).
17. Carter, Y., Sippel, R. S. & Chen, H. Hypothyroidism After a Cancer Diagnosis: Etiology, Diagnosis, Complications, and Management. *Oncologist* **19**, 34–43 (2014).
18. Jiang, P., Du, W. & Wu, M. Regulation of the pentose phosphate pathway in cancer. *Protein Cell* **5**, 592–602 (2014).
19. Kowalik, M. A., Columbano, A. & Perra, A. Emerging role of the pentose phosphate pathway in hepatocellular carcinoma. *Front. Oncol.* **7**, 1–11 (2017).
20. Xiao, J. & Chen, H. S. Biological functions of melanoma-associated antigens. *World J. Gastroenterol.* **10**, 1849–1853 (2004).
21. UCSC. Genome Browser. Available at: <https://genome.ucsc.edu/>. (Accessed: 15th June 2017)