

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| ☐ | ☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☐ | ☒ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	No software was used to collect data, it was downloaded directly from the data sources listed.
Data analysis	All analyses were performed in the Python 3.10 and R 3.6.1 environments. Data analysis was conducted using Pandas 1.5.3, SciPy 1.10.0, Pingouin 0.5.3, and Statsmodels 0.13.5. Data were visualized with Seaborn 0.12.1 and Matplotlib 3.7.1. No statistical method was used to predetermine sample size, the experiments were not randomized, and the Investigators were not blinded to allocation during experiments and outcome assessment. Samples with a mean methylation fraction (across all CpG sites) over three standard deviations from the mean methylation fraction of all samples together were excluded from the analyses. Specific statistical approaches used are noted in the respective methods sections and figure captions.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

All data analyzed was from The Cancer Genome Atlas PAN-CAN cohort^{45–47} (<http://xena.ucsc.edu/>) and the Pan-Cancer Analysis of Whole Genomes⁴⁸ (<https://dcc.icgc.org/pcawg/>). Data can be accessed from the provided links and is described further in the respective publications (<https://doi.org/10.1038/ng.2764>, <https://doi.org/10.1038/s41586-020-1969-6/>). Data to replicate the figures in this manuscript can be found on the "Somatic mutation as an explanation for epigenetic aging (Koch et al. 2024)"⁹⁴ figshare (https://figshare.com/projects/Somatic_mutation_as_an_explanation_for_epigenetic_aging_Koch_et_al_2024_/224232). The panel of normal and gnomAD resources used for filtering the somatic mutation calls can be accessed by downloading Mutect2 (<https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2>). A file containing Illumina 450k array CpG locations and characteristics can be accessed on the Illumina website (https://webdata.illumina.com/downloads/productfiles/humanmethylation450/humanmethylation450_15017482_v1-2.csv). The hg19 genome annotation can be accessed through the University of California Santa Cruz website (<https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/cpgIslandExt.txt.gz>).

Research involving human participants, their data, or biological material

Policy information about studies with [human participants or human data](#). See also policy information about [sex, gender \(identity/presentation\), and sexual orientation](#) and [race, ethnicity and racism](#).

Reporting on sex and gender	Gender was based on self-reporting. Analyses of mutation-associated changes in methylation were carried out comparing individuals of the same gender as there are known sex differences in methylation.
Reporting on race, ethnicity, or other socially relevant groupings	Race was not considered in any analyses.
Population characteristics	The population spanned 14-90 years of age with an average of 58.7 years (standard deviation 14.6). Gender distribution was highly dependent on cancer type, but across all individuals 51.2% identified as male and 48.8% as female.
Recruitment	Patients were recruited by TCGA and PCAWG consortia with criteria that can be found in the respective publications (https://doi.org/10.1038/ng.2764 , https://doi.org/10.1038/s41586-020-1969-6/).
Ethics oversight	Ethics approval was garnered by the TCGA and PCAWG consortia (https://doi.org/10.1038/ng.2764 , https://doi.org/10.1038/s41586-020-1969-6/).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The same size was 9,331 human individuals. This was based on the availability of public data consisting of both somatic mutation and DNA methylation.
Data exclusions	Samples with a mean methylation fraction (across all CpG sites) over three standard deviations from the mean methylation fraction of all samples together were excluded from the analyses.
Replication	A large sample size, nested cross-validation, and replication in normal tissues was used to ensure reproducibility.
Randomization	Experimental groups were defined based on biological characteristics therefore randomization was not relevant.
Blinding	Blinding was not relevant as this was an observational study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks

Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.

Novel plant genotypes

Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.

Authentication

Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosaicism, off-target gene editing) were examined.