UNIVERSITY OF CALIFORNIA SAN DIEGO

Genomic and Epigenomic Aging: Mechanisms and Therapies

A Dissertation submitted in partial satisfaction of the requirements
for the degree Doctor of Philosophy

in

Bioinformatics and Systems Biology

by

Zane Koch

Committee in charge:

Professor Trey Ideker, Chair
Professor Prashant Mali, Co-Chair
Professor Ludmil Alexandrov
Professor Eran Mukamel
Professor Wei Wang

2026

The Dissertation of Zane Koch is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2026

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

Chapter 1, in full, is a reprint of the material as it appears in Nature Aging 2025. Koch, Zane; Li, Adam; Evans, Daniel S.; Cummings, Steven; Ideker, Trey. The dissertation author was the primary investigator and author of this paper.

Chapter 2 contains unpublished material coauthored with Koch, Zane; Li, Adam.; and Ideker, Trey. The dissertation author was the primary author of this chapter.

Chapter 3 contains unpublished material coauthored with Koch, Zane; Nandi, Shuvro P.; Licon, Kate; Bujarrabal-Dueso, Arturo; Meyer, David H.; Saeed, Safa; Perampalam, Pirunthan; Dick, Frederick A.; Schumacher, Björn; Alexandrov, Ludmil B.; Ideker, Trey. The dissertation author was the primary author of this chapter.

VITA

2021 Bachelor of Science in Computational and Systems Biology, University of California Los Angeles

2026 Doctor of Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

Genomic and Epigenomic Aging: Mechanisms and Therapies

by

Zane Koch

Doctor of Philosophy in Bioinformatics and Systems Biology

University of California San Diego, 2026

Professor Trey Ideker, Chair

Professor Prashant Mali, Co-Chair

Aging is accompanied by widespread molecular changes, yet the relationships among these changes – and their ultimate causes – remain incompletely understood. This dissertation

investigates the epigenetic and genomic alterations that accumulate during mammalian aging, their mechanistic origins, and potential strategies for intervention.

First, I characterize the landscape of epigenetic aging across multiple layers of the epigenome. Through comprehensive analysis of six histone modifications and DNA methylation in >1,000 humans and mice, I demonstrate that age-related epigenetic changes are not isolated phenomena but rather reflect a coordinated remodeling process. Epigenetic changes across all layers converge upon a common set of genes, enabling the construction of a "pan-epigenetic" clock capable of predicting age from any epigenetic layer across species.

Second, I establish that somatic mutations and epigenetic change are intimately coupled: mutations coincide with extensive remodeling of the surrounding methylome, and mutation-based age predictions mirror those derived from DNA methylation. These findings suggest that the progressive accumulation of somatic mutations may drive the epigenetic drift observed during aging.

Third, I identify the DREAM complex (Dp, Rb-like-1, E2f, And MuvB) as a key regulator of somatic mutation and aging. DREAM transcriptionally represses DNA repair pathways, and I show that DREAM activity predicts somatic mutation rates, lifespan across 92 mammalian species, and Alzheimer's disease risk. Notably, genetic disruption of DREAM in mice reduces mutation accumulation *in vivo*.

Together, these studies connect epigenetic aging, somatic mutation, and the transcriptional regulation of DNA repair – offering new perspectives on both the mechanisms of aging and opportunities for intervention.

Chapter 1 ADAVANCE SOMATIC MUTATION AS A CAUSE OF EPIGENETIC AGING

Originally published in Nature Aging, January 2025

nature aging

Analysis

https://doi.org/10.1038/s43587-024-00794-x

Somatic mutation as an explanation for epigenetic aging

Received: 8 December 2023

Accepted: 12 December 2024

Published online: 13 January 2025

Check for updates

Zane Koch[1], Adam Li[1], Daniel S. Evans[2,3], Steven Cummings[2,3] & Trey Ideker[1,4]

DNA methylation marks have recently been used to build models known as epigenetic clocks, which predict calendar age. As methylation of cytosine promotes C-to-T mutations, we hypothesized that the methylation changes observed with age should reflect the accrual of somatic mutations, and the two should yield analogous aging estimates. In an analysis of multimodal data from 9,331 human individuals, we found that CpG mutations indeed coincide with changes in methylation, not only at the mutated site but with pervasive remodeling of the methylome out to ±10 kilobases. This one-to-many mapping allows mutation-based predictions of age that agree with epigenetic clocks, including which individuals are aging more rapidly or slowly than expected. Moreover, genomic loci where mutations accumulate with age also tend to have methylation patterns that are especially predictive of age. These results suggest a close coupling between the accumulation of sporadic somatic mutations and the widespread changes in methylation observed over the course of life.

Practically since the elucidation of the DNA double helix, it has been postulated that progressive damage to this fundamental structure is the cause of aging[1,2]. The primary support for this theory relates to somatic mutations, which accumulate in the genomes of most tissues and species throughout life[2–4]. Such accumulation has been associated with multiple characteristics of old age, including immune dysfunction[5], neurodegeneration[6,7] and cancer[8].

Aging has also been associated with other major types of molecular changes beyond DNA mutations, leading to a debate on which of these aging 'hallmarks' are fundamental causes[9,10]. In particular, much recent attention has been given to associations of age with DNA methylation, a dynamic epigenetic mark found primarily at CG dinucleotides (CpG sites) throughout the genome[11]. CpG methylation has diverse functional consequences, including X chromosome inactivation[12], chromatin and transcriptional regulation[13], cell-type specification and maintenance of pluripotency[14–16]. DNA methylation patterns have been found to change regularly over the course of life, prompting the creation of statistical models termed 'epigenetic clocks', which attempt to predict an individual's age using their DNA methylation profile[17,18]. Subsequent research has shown that epigenetic clock predictions correlate with a host of age-related biological attributes, including frailty, Alzheimer's disease, all-cause mortality and time to death[19–22]. Such observations have bolstered epigenetic theories of aging, which propose that progressive remodeling of the epigenome leads to aging phenotypes through the dysregulation of gene expression, cellular function and senescence[23–25]. However, the degree to which epigenetic changes are direct causes of aging remains unclear.

Despite the separate interest in DNA mutations and DNA methylation as theories of aging, the relationship between the two processes is not well understood. One recent study reported that somatic mutations in DNA-binding sites for ten–eleven translocation methylcytosine dioxygenase 1 (TET1), the primary enzyme involved in the removal of methylation marks[26], are associated with local hypermethylation[27]. Another demonstrated an association among somatic mutations, subsequent clonal expansion of blood cells and accelerated epigenetic aging[28]. Most other research linking DNA sequence and methylation has focused on inherited germline variants rather than acquired somatic mutations, such as efforts to identify methyl quantitative trait loci

[1]Program in Bioinformatics and Systems Biology, University of California, San Diego, La Jolla, CA, USA. [2]California Pacific Medical Center Research Institute, San Francisco, CA, USA. [3]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA, USA. [4]Department of Medicine, University of California, San Diego, La Jolla, CA, USA. e-mail: steven.cummings@ucsf.edu; tideker@health.ucsd.edu

linking common polymorphisms to methylation levels of specific CpG sites[29,30].

An intrinsic biochemical connection between DNA mutation and methylation occurs at 5-methylcytosine residues[3,31,32], which spontaneously deaminate over time to yield thymine[33]. A prerequisite for this mutational event is cytosine methylation, relating somatic mutation to prior epigenetic modification of DNA. Conversely, a prerequisite for DNA methylation is the presence of a CpG site, which may be eliminated by prior somatic mutation. Given this interdependence, we considered that the separate links that have been established between DNA mutation and aging and between DNA methylation and aging might each reflect a common underlying process whereby methylation potentiates mutation and/or mutation potentiates changes in methylation.

To explore this hypothesis, we set out to comprehensively examine the relationship between somatic DNA mutations and DNA methylation in large collections of human tissue samples characterized for both layers of molecular information. In what follows, we identify several types of interaction between somatic mutation and DNA methylation, both one-to-one and one-to-many (Extended Data Fig. 1a). Based on these findings, we use somatic mutations as a surrogate for epigenetic marks in measures of aging, indicating the degree to which epigenetic aging is explained by somatic mutations (Extended Data Fig. 1b). These results are initially demonstrated in cancer cohorts, which have by far the greatest numbers of individuals with matched DNA mutation and methylation profiles. We then show that our findings extrapolate to normal tissues drawn from the same patients.

Results

Genome-wide hypomethylation of mutated CpG sites

To study the connections between somatic mutations and DNA methylation marks, we first analyzed data from human patients cataloged in The Cancer Genome Atlas (TCGA)[34–36] and the Pan-Cancer Analysis of Whole Genomes (PCAWG)[37]. Tumor biopsies had been drawn from a diversity of tissue types and characterized by whole-exome sequencing (TCGA, 8,680 exomes across 33 tissues) or whole-genome sequencing (PCAWG, 651 genomes across 3 tissues). In each case, DNA from the tumor sample was compared to a second DNA sample drawn from the same individual, with differences used to define somatic mutations (typically comparing the tumor DNA sequence to that in whole blood; Methods). These data had been complemented by methylation profiling of the same tissues using the Illumina Infinium HumanMethylation450 BeadChip, which provides methylation fraction readouts (the fraction of DNA reads that are methylated) for approximately 450,000 CpG sites across the entire genome[38]. To facilitate the validation of our findings in a nontumor context, we additionally obtained somatic mutation ($n = 111$) and DNA methylation ($n = 187$) profiles from normal, noncancerous tissues from a subset of these same individuals (Methods).

From these data, we considered all single base-pair substitution mutations ($n = 3{,}457{,}875$ mutation events) and CpG sites for which all individuals had a reliably measured methylation value ($n = 326{,}751$ CpG sites; Methods and Extended Data Fig. 2a–d). Consistent with previous reports[3,31], CpG sites were the most frequently mutated dinucleotide, accounting for 13.5% of all somatic mutations across the genome (Fig. 1a). The vast majority of these were C > T transitions (82.3%; Extended Data Fig. 2e) occurring at CpG sites that tended to be more heavily methylated than average (Extended Data Fig. 2f).

We next asked whether individuals harboring a mutated CpG site exhibit lower detected levels of methylation at that site compared to nonmutated individuals (Fig. 1b). We reasoned that once mutated, the site would no longer constitute a CG dinucleotide, reducing its likelihood of methylation. Indeed, we found a significant decrease in detected methylation in individuals with a mutation at a CpG site compared to nonmutated individuals at the same site (Mann–Whitney $P = 3.90 \times 10^{-9}$; Fig. 1c and Methods), with the decrease of detected methylation related to the frequency of reads with the mutant allele (Pearson's $r = -0.17$, $P = 2.08 \times 10^{-53}$; Fig. 1d and Extended Data Fig. 2g). These results support a model in which CpG mutations occur primarily at hypermethylated sites (due to the spontaneous deamination of methylcytosine) and can become fixed in the genome of daughter cells, causing a decrease in observed methylation corresponding to the mutated clonal population.

Mutated sites show extensive remodeling of the surrounding methylome

During this exploration, we noted numerous cases in which somatic mutations coincided not only with hypomethylation at the mutated CpG site but also with atypical methylation of numerous CpGs in the surrounding genome. An illustrative example was the C > T mutation at base pair 56,642,556 of chromosome 16 in the individual TCGA-GV-A3QI (Fig. 2a). CpG sites adjacent to this somatic mutation were strikingly hypermethylated in this individual, with such hypermethylation extending over a contiguous region more than 30 kilobases (kb) downstream. This effect encompassed the metallothionein 2A gene (*MT2A*) as well as additional metallothionein family members (*MT1E* and *MT1M*), for which methylation-linked repression has been associated with metastasis in multiple cancer types[39–42].

To move beyond anecdotal observations, we devised a general test for whether somatic mutations are associated with remodeled methylation at surrounding CpGs. We computed a quantity we called ΔMF_w, the median change in methylation fraction observed for CpGs within a window of $\pm w$ kb near a mutated site, comparing the mutated individual to matched nonmutated individuals (Fig. 2b and Methods). Examination of $\Delta MF_{1\,kb}$ values at increasing distances from mutated sites revealed significantly greater methylation change at these sites compared to randomly chosen nonmutated sites, with this effect extending for up to ±10 kb (Fig. 3a and Methods). Calculating the methylation change across this entire ±10-kb range, we observed that $\Delta MF_{10\,kb}$ tended toward substantially more extreme values than expected at random ($n = 2{,}600$ mutated sites with sufficient nearby CpGs, $P < 10^{-124}$; Fig. 3b), with mutated loci more than four times as likely to have an extreme decrease in nearby methylation ($\Delta MF_{10\,kb}$ of <−0.3; Fig. 3c). Mutated loci were also enriched for nearby methylation increases (similar to the example in Fig. 2a), albeit more weakly (Fig. 3c). This phenomenon was observed in all tissue types for which we had data (Extended Data Fig. 3a). Overall, 15.5% of mutated sites had a more extremely positive or negative $\Delta MF_{10\,kb}$ than expected (against a reference of random control sites, $|z\ \text{score}| > 1.96$; Methods).

Deeper explorations showed that the methylation increases and decreases (1) have a direction of change that depends on local CpG density (that is, whether they are inside CpG islands[13]; Fig. 3d and Extended Data Fig. 3d), (2) are specific to the genomic context, enriched for being mutations to CpG sites (Fig. 3d) in intergenic regions (Extended Data Fig. 3e,f), and (3) increase with the fraction of DNA in the sample harboring the mutation (Fig. 3e and Extended Data Fig. 3g). In noncancerous tissues, methylation disturbances were found surrounding 8.0% of somatic mutations ($|z\ \text{score}| > 1.96$ compared to random control sites; Methods and Extended Data Fig. 4a), each extending for, on average, ±1 kb from the site of mutation (Extended Data Fig. 4b). This decreased frequency and shorter range of effect may be attributed to the lower mutated allele frequency (MAF) of somatic mutations in normal tissues compared to tumor tissues (mean ± s.d. MAF: normal = 0.10 ± 0.013, tumor = 0.83 ± 0.24). These results indicate that our earlier observations were not anecdotal, but CpG mutations are generally associated with an atypical methylation pattern in the surrounding DNA.

Somatic mutations mirror epigenetic predictions of age

While mutation of any particular CpG site is exceedingly rare in the human population, and thus a poor predictor of age, its corresponding CpG methylation fraction varies regularly in a manner often associated

https://doi.org/10.1038/s43587-024-00794-x

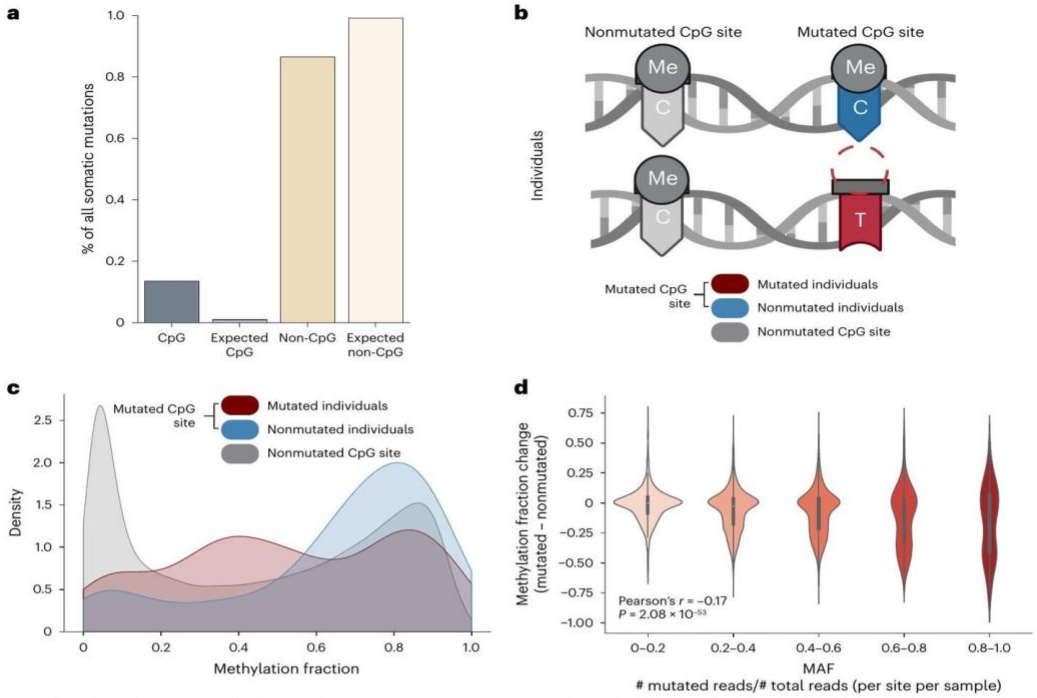


Fig. 1 | Frequency and methylation status of CpG mutation events. **a**, Percentage of genome-wide somatic mutations classified as CpG (n = 467,079 mutations) or non-CpG (n = 2,990,796 mutations). Expected percentages were calculated supposing mutation probability to be uniform across the genome (Methods). **b**, Diagram showing two categories of CpG sites: those where no individual is mutated (nonmutated CpG site, gray) and those where a mutation has occurred in at least one individual (mutated CpG site, red; bottom) and the remaining individuals are nonmutated (blue; top). **c**, Distribution of CpG methylation values for the categories of CpG sites from **b**. The methylation fractions of mutated individuals (red) and nonmutated individuals (blue) are shown for the 1,000 CpG sites with the highest MAF (corresponding to MAF > 0.53; Methods). **d**, Methylation change between mutated and nonmutated individuals at n = 8,037 mutated CpG sites. Methylation change is the difference between the median methylation fraction in mutated individuals and the median methylation fraction in nonmutated individuals of matched age and tissue. CpG sites are binned into five groups based on MAF, with violin plots summarizing the distribution of methylation changes within each group. Vertical bars inside each violin represent the interquartile range. The two-sided P value was calculated based on the exact distribution of Pearson's r modeled as a beta function. MAF, mutant allele fraction.

with age[17]. However, we considered that the one-to-many relationship revealed by our previous analysis (Fig. 3), by which a single CpG mutation maps to a broad profile of methylation changes in the surrounding DNA, might bridge this apparent gap between sporadic mutation accumulation and consistent methylation change. Accordingly, we compared two procedures for predicting human chronological age: the first using an individual's profile of CpG methylation values, as in previous epigenetic aging models (methylation clock), and the second using their profile of somatic mutations, including the counts of somatic mutations within 10 kb of each of these same CpGs (mutation clock; Methods and Fig. 4a). Evaluating these models using a nested cross-validation procedure (Methods), we found that the methylation clock predicted age with an accuracy of r = 0.83 (Pearson's correlation), whereas the mutation clock had an accuracy of r = 0.67 (Fig. 4b,c). When predictions were examined within each tissue, both the mutation and methylation models were most accurate at predicting age in brain samples and least accurate in kidney and bone samples (Extended Data Fig. 5a).

Beyond their accuracies of age prediction, we found that the two clocks agreed significantly in several other key aspects. First, the predictions from both models were highly correlated across individuals (Pearson's r = 0.74), and this relationship persisted even after controlling for calendar age (partial correlation = 0.45, $P = 1.48 \times 10^{-82}$; Fig. 4d and Methods). For example, for individuals predicted by mutations to be 1 year older than their calendar age, the methylation clock yielded a corresponding overprediction of 0.75 ± 0.53 years (mean ± s.d.). This same agreement in over-/underprediction of each individual (similarity in model residuals) was observed when comparing the mutation clock to previously published methylation clocks (Fig. 4e and Extended Data Fig. 6a,b)[17,18,21]. Second, CpG sites for which the surrounding mutation burden was most associated with age also tended to have the most age-associated methylation values, as highlighted by the 80-fold greater-than-expected overlap among the top 1% of age-associated sites of each kind (Fig. 4f and Methods). One example was CpG site cg19236454 (chr19:42,799,926), for which the local mutation burden progressively increased with age (±10 kb, r = 0.18), whereas the methylation of this site was progressively lost (r = −0.18; Fig. 4g). This overlap demonstrates a local coupling of age-related mutation accumulation and methylation change, independent of any clock or model.

To confirm that the synchronization between the mutation and methylation clocks is not unique to tumor tissues, we applied the same modeling procedure to normal tissues (n = 111 somatic mutation profiles, n = 187 methylation profiles). We found that models trained to predict chronological age from the mutation and methylation profiles of these normal tissues (1) were predictive of age (Pearson's

https://doi.org/10.1038/s43587-024-00794-x

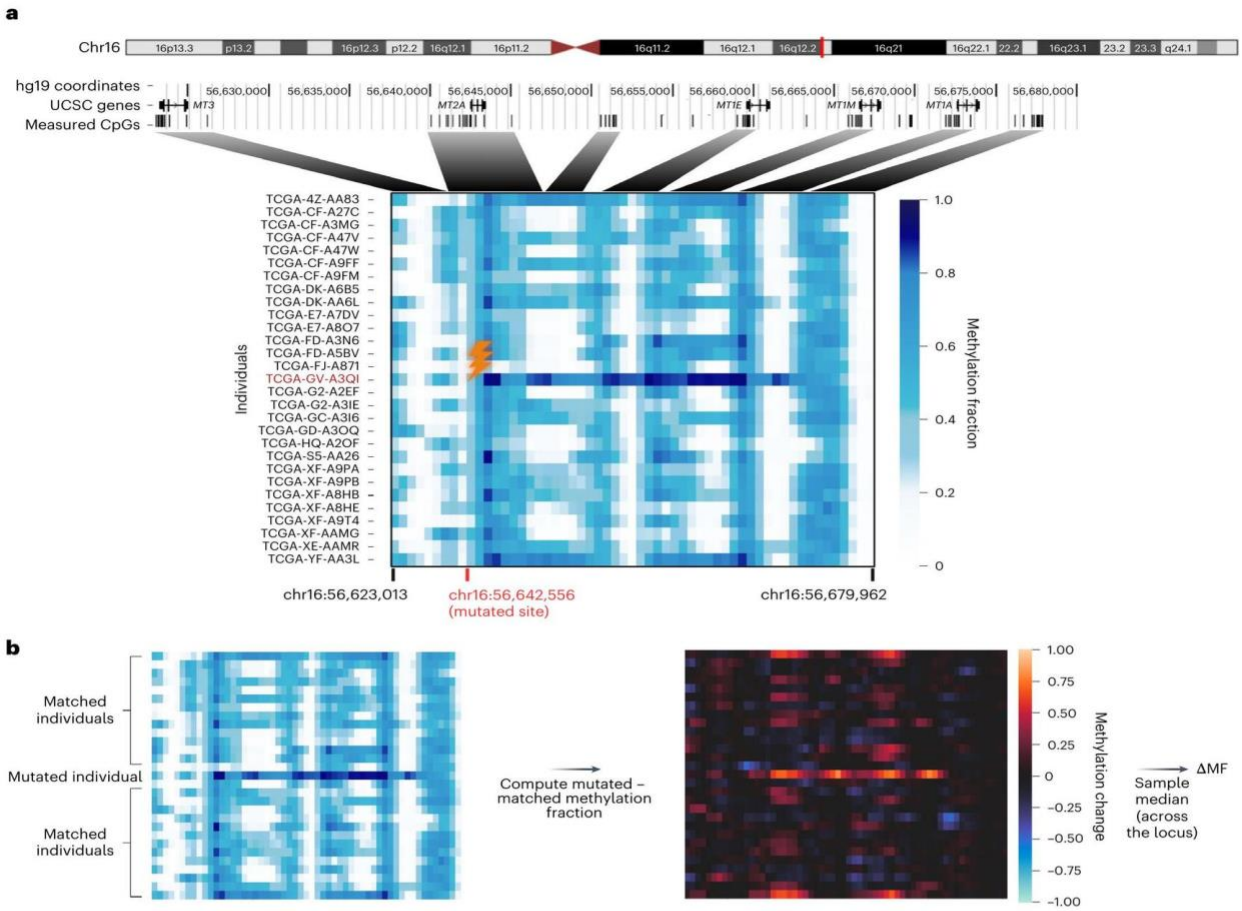


Fig. 2 | Association of mutations with regional methylation patterns. **a**, Example mutated site where the individual TCGA-GV-A3QI has a C > T mutation at chr16:56,642,556 of the hg19 human genome. Top, ideogram of chromosome 16, with a red bar indicating the location of the mutated site. The first underlying track shows hg19 base-pair coordinates, the second the documented genes in the region (encoding five metallothionein factors) and the third the locations of CpG sites measured on the Illumina 450k methylation array (vertical bars). Bottom, heatmap of CpG methylation fractions. Rows are samples (1 mutated, 28 matched), and columns are the measured CpGs within a ±50-kb window proximal to the mutation (n = 62 CpG sites). Color corresponds to the methylation fraction of each CpG. The mutated sample row and mutated site column are labeled in red, with the mutation event indicated by a lightning bolt. **b**, Calculation of the change in methylation fraction (ΔMF) with reference to a specific mutated site. Left, heatmap of methylation fractions of the mutated site and CpGs in the surrounding window, replicated from **a**. Right, heatmap of the corresponding differences in methylation between each sample (row) and all other samples in the matrix (median of other rows), computed separately for each site in the window (columns). The final ΔMF value was calculated as the overall methylation change of the mutated sample, taking the median across all sites in the window (Methods). Matched background samples were defined as those without any somatic mutations in the window and that were of the same tissue type and approximate age (±5 years) as the mutated sample. UCSC, University of California, Santa Cruz.

r: mutation clock = 0.74, methylation clock = 0.92; Extended Data Fig. 5b), (2) agreed in the age prediction of the same individuals as clocks trained on tumor data (Pearson's r of normal versus tumor predicted age: mutation clocks = 0.79, methylation clocks = 0.89; Extended Data Fig. 5c,d) and (3) continued to exhibit an association between mutation and methylation age (partial correlation of normal clocks = 0.48, P = 0.0018; Methods and Extended Data Fig. 5e–g). Thus, mutation and methylation profiles were synchronized with respect to predictions of age (Fig. 4h), both globally (throughout the genome) and locally surrounding individual CpG sites, in normal and tumor tissues.

Discussion

In this study, we observed notable associations between CpG mutation and methylation at multiple scales. At the scale of single nucleotides, CpG sites altered by somatic mutation (the most frequent mutation type across the genome; Fig. 1a and Extended Data Fig. 2e,f) exhibit a decrease in detected methylation at that site (Fig. 1c,d). At a larger scale, such mutated sites coincide with sweeping methylation changes across numerous CpGs within the surrounding genomic region (Figs. 2 and 3a–c). Plausibly as a result of this larger-scale relationship, individuals whose mutations indicate increased genomic age also tend to have older methylomes (Fig. 4d,h and Extended Data Fig. 5g).

https://doi.org/10.1038/s43587-024-00794-x

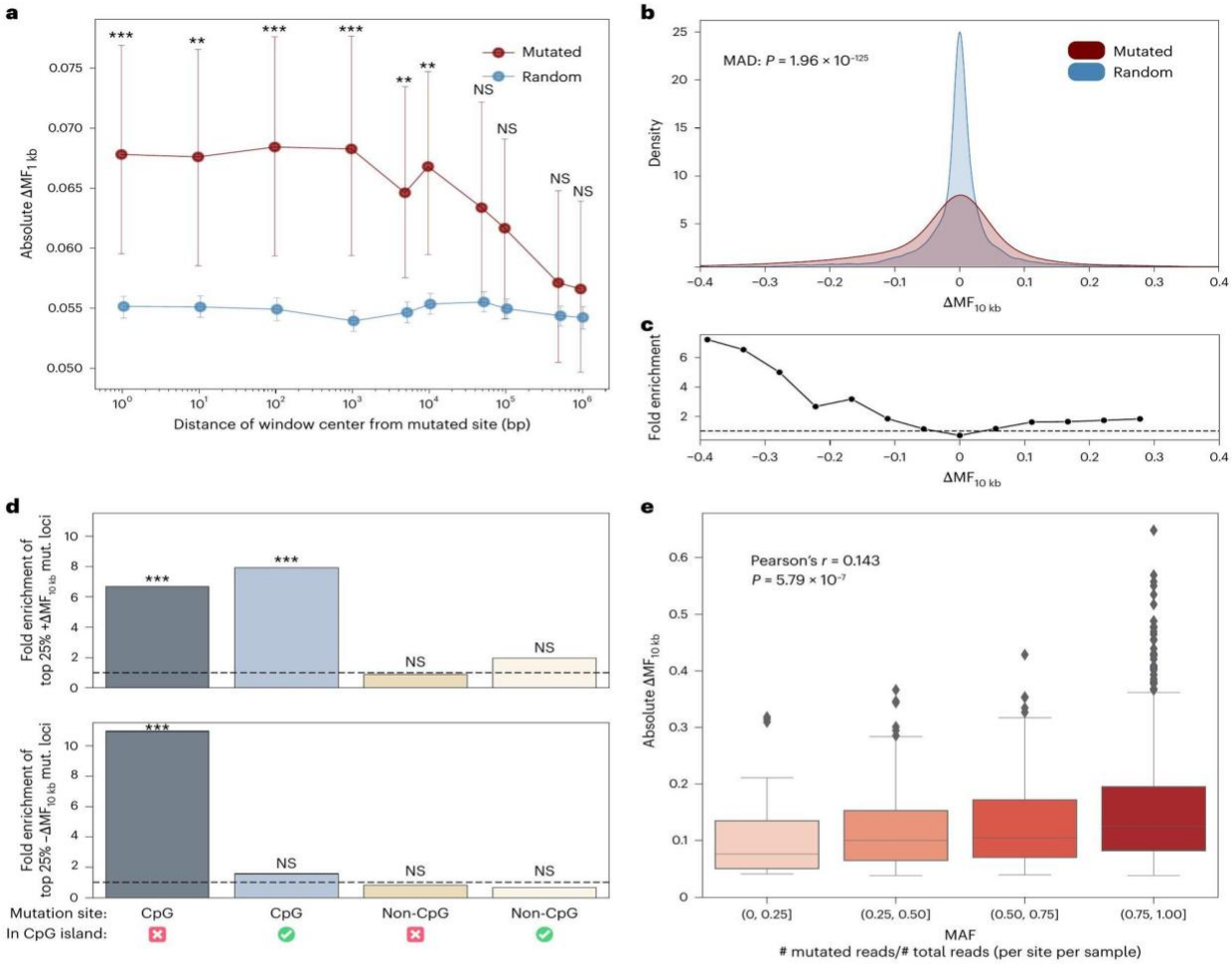


Fig. 3 | Magnitude and extent of methylation changes near somatic mutations. **a**, Absolute $\Delta MF_{1\,kb}$ as the window center is moved away from the mutated site ($n = 2{,}600$, red). This quantity is also shown for nonmutated random control sites ($n = 260{,}000$, blue) (Methods). Points indicate the mean value, and error bars denote the 95% confidence interval. A significant difference in the distribution of absolute $\Delta MF_{1\,kb}$ values (two-sided t test) is marked (***$P \leq 0.001$, **$P \leq 0.01$). Other comparisons are nonsignificant (NS, $P > 0.05$). **b**, Probability distribution of $\Delta MF_{10\,kb}$ values calculated in a ±10-kb window surrounding mutated (red) versus random control (blue) sites. Mutated sites include $n = 2{,}600$ mutated sites with MAF ≥ 0.8, ≥15 matched individuals (individuals of the same tissue type within ±5 years of age) and one or more measured CpGs within the window. Random control sites include $n = 260{,}000$ nonmutated sites (Methods). The same was found when controlling for the initial methylation state of mutated and random loci (Extended Data Fig. 3b,c). The P value is shown for a two-sided Mann–Whitney test for a difference in median absolute deviation (MAD) of $\Delta MF_{10\,kb}$ between the mutated and nonmutated random control loci. **c**, Line plot depicting the fold enrichment for mutated over nonmutated sites as a function of $\Delta MF_{10\,kb}$. Fold enrichment is the ratio of the probability of observing a given $\Delta MF_{10\,kb}$ for mutated sites versus the probability of that $\Delta MF_{10\,kb}$ for nonmutated control sites. $\Delta MF_{10\,kb}$ is divided into equally spaced bins from −0.4 to 0.4. **d**, Enrichment of extreme $\Delta MF_{10\,kb}$ values at CpG sites and CpG islands. Top versus bottom bar charts show the 25% of mutations with the most positive versus most negative $\Delta MF_{10\,kb}$ values in **a** ($n = 650$ mutations each). The enrichment of these mutations (bars, y axis) was considered for different types of sites, depending on whether the site is a CpG and/or falls within a CpG island (x-axis categories). Enrichment was compared to the genomic baseline (Methods), with significance determined by a one-sided binomial test. Significant enrichment ($P \leq 0.001$) is marked with asterisks (***) and nonsignificant ($P > 0.01$) with 'NS'. CpG islands are defined as genomic regions with ≥200 bp, ≥50% GC content and a high CpG occurrence. **e**, Boxplot of the absolute $\Delta MF_{10\,kb}$ value as a function of the MAF. The plot includes all mutated sites with ≥15 matched samples and one or more measured CpGs within ±10 kb ($n = 3{,}880$ mutated loci). The two-sided P value was calculated based on the exact distribution of Pearson's r modeled as a beta function. The central line represents the median; the edges of the box represent the interquartile range; the whiskers indicate 1.5 times the interquartile range; and the points represent all $\Delta MF_{10\,kb}$ values outside these ranges.

A fundamental tension addressed in this study is that two individuals rarely share a somatic mutation at the same CpG site; thus, mutations would initially seem too sparse to explain the numerous CpG sites at which methylation reliably changes with age. However, our findings show that single mutations can correspond to appreciable shifts in the methylome, with a graded relationship that depends on the frequency of the mutated allele (that is, clonality of the mutant cell population; Fig. 3e). Consistent with these findings, we see that, within

https://doi.org/10.1038/s43587-024-00794-x

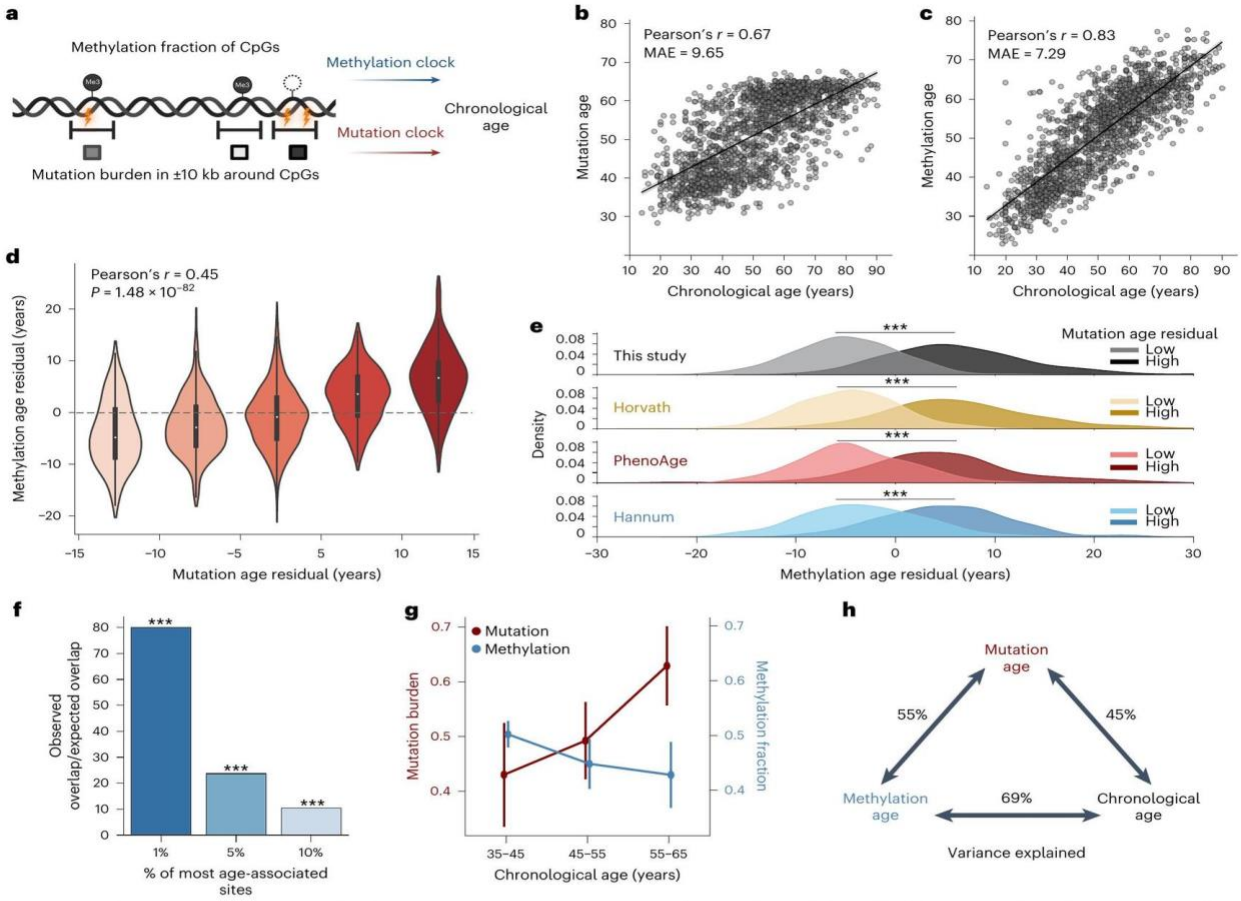


Fig. 4 | Association among mutation age, methylation age and chronological age. a, Methylation clock: the methylation fractions of CpGs are used in a gradient boosted tree model to predict chronological age. Mutation clock: the count of mutations around the same CpGs are used in an identical model to predict chronological age. Both models incorporate similar covariates and whole-genome features (Methods). **b**, Scatter plot of human individuals, showing age predictions from the mutation model versus their chronological age. The plot includes $n = 1{,}601$ individuals with samples from five tissues (Methods). **c**, Similar to **b** but showing age predictions from the methylation rather than mutation model for the same individuals. **d**, Violin plots of the methylation age residual versus the mutation age residual (Methods). The plot includes the same individuals as in **b** and **c**. Pearson's *r* refers to the correlation between the methylation age residual and the mutation age residual (that is, partial correlation, $P = 1.48 \times 10^{-82}$, two-sided *P* value calculated based on the exact distribution of Pearson's *r* modeled as a beta function). The central line of the inner boxplot represents the median; the edges of the box represent the interquartile range; and the whiskers represent 1.5 times the interquartile range. **e**, Distribution of methylation age residuals for the same individuals as in **b** and **c**, computed according to each of four previous methylation clocks. 'This study' refers to the methylation clock shown in **c** (Methods). For each clock, the 20% ($n = 320$) of individuals with the youngest mutation age for their chronological age are shown in a lighter color (low mutation age residual), and the 20% ($n = 320$) of individuals with the oldest mutation age for their chronological age are shown in a darker color (high mutation age residual). The asterisks (***) indicate a significant ($P \leq 10^{-51}$) difference in distribution between the low and high mutation residual age groups, based on a two-sided Mann–Whitney *U* test. **f**, Bar plot depicting the ratio of observed versus expected overlap between sets of age-associated CpG sites. For the same individuals and CpG sites as in **c**, the CpGs with maximal (top 1%, 5% and 10%) Pearson's correlation between local mutation burden (±10 kb) and age and between methylation fraction and age were chosen. The intersection (overlap) between these sets was compared to the expected intersection assuming random selection (Methods). Significant enrichment based on a two-sided binomial test ($P \leq 10^{-5}$, Bonferroni corrected) is marked with asterisks (***). **g**, Mean mutation burden (left *y* axis) or mean methylation fraction (right *y* axis) plotted versus chronological age (*x* axis) for CpG site cg19236454. Data were from brain (LGG, low grade-glioma) samples, considering individuals with a nonzero mutation burden (±10 kb) at this site ($n = 67$). Pearson's correlation with chronological age: mutation burden = 0.18, methylation = −0.18. Error bars denote the standard error. **h**, Diagram summarizing the relationships among three measures of age: mutation, methylation and chronological time. The variance explained was calculated as the squared Pearson's correlation between each pair of measures for the same individuals as in **b** and **c**. MAE, mean absolute error.

individuals of the same calendar age, mutation and methylation clocks agree on which individuals are aging faster or slower (Fig. 4d,e), and somatic mutations explain more than 50% of the variation in methylation age across individuals (Fig. 4d,h). Notably, this relationship can be observed directly, without relying on predictive modeling, in the colocalization of age-associated mutation and methylation sites (Fig. 4f,g).

The mechanisms by which a CpG mutation affects its methylation state, or conversely by which CpG methylation potentiates its own

mutation, are already established. The prior methylation of a CpG makes a subsequent somatic mutation more likely due to methylcytosine deamination[33]. In turn, when either nucleotide of a CpG site is mutated, the site is no longer a CpG, substantially decreasing the likelihood of future methylation by a DNA methyltransferase[43]. For mutations exhibiting larger-scale gains or losses of methylation in the surrounding kilobases, it is conceivable that either methylation or mutation, or neither, could be the primary causal agent. The observed association between somatic mutation and local hypermethylation (Fig. 3b) may occur if hypermethylation creates an environment prone to methylcytosine deamination events, giving rise to rare somatic mutations embedded within hypermethylated regions. However, this model does not explain the frequently observed co-occurrences of mutations with neighboring hypomethylation throughout the genome (Fig. 3b), as hypomethylation should decrease, not increase, the local probability of mutation.

An alternative possibility is that mutations are the primary causes of subsequent changes in methylation. Mutations within the DNA-binding site of a methylase or demethylase enzyme could plausibly affect enzyme activity, dysregulating the methylation state of the surrounding genome[44–46]. Such a relationship has been reported explicitly for somatic mutations in the DNA-binding sites of TET1, a demethylase, leading to local gains of methylation[27]. More broadly, it is well known that germline DNA sequence variants govern the methylation patterns of many CpG sites, affecting as many as 40% of CpGs throughout the genome (i.e. the effect underlying methyl quantitative trait loci)[47]. Somatic mutations in these sequences may yield effects on methylation analogous to those observed for inherited variants.

A third possibility is that mutation and methylation events are not causal of each other but both are downstream of some earlier event. One such event might relate to the repair of DNA double-strand breaks, which have been demonstrated to result in both somatic mutations and methylation changes near the site of repair[48–50]. Here, the mutation and methylation changes would be indicative of an earlier repair of double-strand breaks, an activity recently suggested to cause epigenetic aging[23]. Similarly, the formation of 8-hydroxyguanine lesions has been associated with both somatic mutation and loss of methylation in the surrounding DNA[51].

Regardless, understanding the causality among mutations, methylation and aging has important implications for how we seek to prevent or reverse aging. In particular, if mutations are the fundamental driver of aging phenotypes, and epigenetic changes simply track this process, then strategies aimed at epigenetic reversal[23,52–54] may be treating a symptom rather than a cause.

Some limitations of this study are as follows. First, most of the samples we analyzed came from tumor biopsies, as the current large human datasets measuring both somatic mutations and methylation pertain to cancer patients. While analyses of normal tissues from a subset of these same individuals supported the findings in tumor tissues (Extended Data Figs. 5 and 7), the relevance of these findings to normative aging should be further examined in larger datasets of normal individuals and tissues by, for example, applying mutational clocks to independent datasets. This limitation notwithstanding, we note that mutation burdens in cancerous and normal tissues are similar[55,56], and most mutations found in tumors are believed to represent normal mutational processes unrelated to cancer[55,57]. Second, most of the somatic mutations we analyzed were derived from exome sequencing, limiting our conclusions regarding the association of mutations with age and methylation to those regions. Third, our analysis was cross-sectional rather than longitudinal, with each individual measured at a single time point only. In the future, a longitudinal study design could greatly inform the actual order of events. Fourth, methylation values were assessed using Illumina 450k technology, which scores the methylation fractions of approximately 450,000 defined CpG sites through bisulfite conversion of unmethylated cytosines (to thymines) and subsequent oligonucleotide hybridization[58]. For CpGs at which the cytosine has been mutated (for example, to TpG), the technology will read out this mutation event as a loss of methylation, whether or not a concomitant loss of methylation has occurred. Finally, there exist other factors associated with epigenetic aging that do not explicitly implicate somatic mutations. Some epigenetic changes clearly reflect alterations in tissue composition with age[59,60], and other changes are associated with the expression of developmental genes[61,62], such as in the binding sites of the polycomb repressive complex[63,64]. Some of these factors may nonetheless relate to DNA mutations; for instance, somatic mutations can drive alterations in tissue composition[65,66].

Methods

Data access and preprocessing

We obtained paired DNA methylation (Illumina 450k array) and somatic mutation data from two public consortia: TCGA[34–36] and PCAWG[37]. Relevant to TCGA, we used the Pan-Can cohort (http://xena.ucsc.edu/), which includes 8,680 samples from 33 cancer types with both Illumina 450k methylation data and somatic mutation calls. In addition to these tumor samples, we generated somatic mutation profiles from normal, noncancerous tissues of 111 TCGA individuals (see 'Somatic mutation identification' in the Methods) and obtained DNA methylation profiles from normal tissues of 187 TCGA individuals (40 individuals had both normal somatic mutation and normal DNA methylation profiles). Data from the PCAWG consortium (https://xenabrowser.net/datapages/?hub=https://pcawg.xenahubs.net:443) include 651 samples from three cancer types with both Illumina 450k methylation array data and whole-genome somatic mutation calls. Methylation data from both cohorts were further processed as follows. First, we removed CpG sites for which any sample had a missing value, leaving 273,202 CpG sites for TCGA and 326,749 CpG sites for PCAWG (Extended Data Fig. 2a–d). Second, we removed samples for which the mean methylation fraction (over all remaining CpGs) was more than 3 s.d. outside its expected (mean) value over all samples. Third, each sample was quantile normalized.

Somatic mutation identification

In addition to the somatic mutations called by the TCGA consortium in tumor tissues[35] (n = 8,860 individuals), we identified somatic mutations in the normal tissues of individuals from TCGA who had DNA sequencing carried out in three locations: tumor, tumor-adjacent tissue and blood (n = 111). We called somatic mutations using Mutect2 (ref. 67) by comparing the tumor-adjacent DNA sequence to the blood DNA sequence in each individual. Mutect2 was run with default parameters following the tumor versus normal somatic mutation calling pipeline described by the Genome Analysis Toolkit (GATK) but treating the tumor-adjacent DNA sequence as the 'tumor' sample and the blood DNA sequence as the 'normal' sample (that is, reference) (https://gatk.broadinstitute.org/hc/en-us/articles/360035531132--How-to-Call-somatic-mutations-using-GATK4-Mutect2). In so doing, we identified variants present in the tumor-adjacent tissue that were not present in the blood of the same individual, indicating their somatic origin. The germline resources provided by the GATK were used to filter out common germline mutations[67]. Mutations at loci with coverage of fewer than 50 reads in either the tumor-adjacent or blood sample were discarded.

Characterizing CpG mutation frequency

Based on the University of California, Santa Cruz, hg19 human genome annotations[68], the number of nucleotides that comprise CpG residues equals 2 bp × 28,299,634 CpG sites, within a total genome length of 3,137,144,693 bp. Therefore, 1.8% of randomly distributed mutations are expected to be CpG mutations, and the remaining 98.2% of mutations are not (Fig. 1a). As CpG sites are palindromic, CG on one DNA strand is equivalent to GC on the complementary strand; thus, for simplicity, we refer to all CpG mutations on either strand as alterations

to the C residue in the first position. This convention was used to record the frequency of each dinucleotide sequence resulting from a CpG mutation (Extended Data Fig. 2e). For these cumulative analyses relating to the overall frequency of CpG mutation (Fig. 1a and Extended Data Fig. 2e,f), the PCAWG samples were used exclusively as they have whole-genome sequences, encompassing all CpG sites, rather than exome sequences only.

Characterizing methylation at mutated CpG sites

The methylation status of two categories of CpG sites was compared: 'nonmutated sites', where no mutation was observed in any individual, and 'mutated sites', where at least one individual had a mutation (Fig. 1b). For CpG sites of the first category (265,399 nonmutated sites), the distribution of methylation fractions was plotted (Fig. 1c). For CpG sites of the second category (8,037 mutated sites), some individuals harbor a mutation at that particular CpG and some do not. In this case, the distributions of CpG methylation fractions were plotted separately for the mutated versus nonmutated individuals (Fig. 1c). For analyses of methylation associated with mutated CpG sites (Fig. 1c,d), tumor TCGA samples were used exclusively as there were many more occurrences of CpG mutations in this dataset due to its much larger sample size. We note that the effect of a CpG mutation on the overall methylation fraction of a tissue can be masked by nonmutated cells of that tissue having higher-than-expected methylation levels—this is a possible explanation for the rare increases in methylation observed at mutated CpG sites and the modest association between MAF and methylation change (Fig. 1d).

Calculating mutation-associated methylation change

Somatic mutation events were defined as site–sample pairs for which the nucleotide present at that site in that particular sample assumed a different (A, C, G, T) value than in the matched normal tissue (typically whole blood from the same sample). For each mutation event, the genomic 'locus' was defined as a $\pm w$-kb window (upstream and downstream) proximal to the mutated site. Matched background samples were defined as those without any somatic mutations in this window and that were of the same tissue type and approximate age (±5 years) as the mutated sample (Figs. 2 and 3). Using these definitions, we calculated $\Delta\mathrm{MF}_w$, the median normalized change in methylation fraction $m_{k,j}$ at each measured CpG site k in the $\pm w$-kb window surrounding mutated site i in mutated sample j relative to matched background samples $b \in$ background:

$$\Delta\mathrm{MF}_{i,j} = \mathrm{median}_{k \in \mathrm{window}(i,w)} \left(m_{k,j} - \mathrm{median}_{b \in \mathrm{background}} \left(m_{k,b}\right)\right)$$

$$\mathrm{Window}\,(i, w) = \{\text{CpG sites within} \pm w \text{ from site } i\}$$

We calculated ΔMF in ±10-kb windows for each mutation event in the PCAWG data, where the mutation was a single base-pair substitution; there were at least 15 viable matched background samples; there was no other mutation within ±10 kb of the mutation in the mutated or matched background samples; the MAF was ≥0.8; and there was at least one CpG site within 10 kb of the mutated base (Fig. 3b,c). For the normal tissue TCGA data, we calculated ΔMF in ±1-kb windows at each mutation event, where the mutation was a single base-pair substitution; there were at least 10 viable matched background samples; there was no other mutation within ±10 kb of the mutation in the mutated or matched background samples; and there was at least one CpG site within 1 kb of the mutated base (Extended Data Fig. 4a–c). These criteria left 2,600 and 463 mutation events for the tumor PCAWG and normal TCGA datasets, respectively. Random control events were chosen to create a background distribution of ΔMF values at genomic loci lacking somatic mutations in any sample. For each true mutation event, we randomly chose 100 nonmutated nucleotides from the corresponding mutated sample and calculated ΔMF at these loci. To perform this calculation, we treated the randomly chosen nucleotide as if it were a mutation and calculated the ΔMF of CpG sites within ±1 or ±10 kb (for normal TCGA or tumor PCAWG, respectively). To control for the initial methylation state, we repeated this analysis by selecting random control sites with a methylation profile matched (in terms of the median methylation fraction of comparison sites) to that of the matched samples of each true mutation event (Extended Data Fig. 3b,c). To quantify the frequency at which mutated sites had ΔMF values that were more extreme than expected, we calculated the z score of the ΔMF values at each mutated site relative to the distribution of ΔMF at all random control sites.

Extent of mutation-associated methylation remodeling

To investigate the extent of methylation remodeling associated with a somatic mutation, we computed ΔMF in a 1-kb window at increasing distances (up to 1 Mb) from each mutated and random control site. Then, we aligned these $\Delta\mathrm{MF}_{1\mathrm{kb}}$ values by their linear genomic distance from their respective mutated site to observe the marginal methylation change at each distance (Fig. 3a and Extended Data Fig. 4c).

CpG and CGI enrichment

The genomic background rate of CpG islands (CGIs) and CpGs was calculated based on the hg19 annotation[69,70] (https://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/cpgIslandExt.txt.gz). The frequency of CpGs in CGIs was based on previously published statistics[31]. To understand whether mutation type (CpG or non-CpG) and location (CGI or non-CGI) were related to the degree of mutation-associated methylation change, we divided the frequency of each mutation type by the expected genome-wide rate ('fold enrichment', focusing on the 25% of mutation events with the most positive or negative $\Delta\mathrm{MF}_{10\mathrm{kb}}$). A one-sided binomial statistic was used to test for an increase in mutation frequency above the genomic background rate of each mutation type (Fig. 3d).

Clock datasets and features

The tumor tissue samples used for all clock-related analyses were from the LGG (low-grade glioma; brain), GBM (glioblastoma multiforme; brain 2), SARC (sarcoma; bone), KIRP (kidney renal papillary cell carcinoma; kidney) or THCA (thyroid carcinoma; thyroid) cancer types in TCGA (n = 1,601 individuals). Samples from normal tissues were chosen from these same tissue types as the tumor samples when available (n: thyroid = 23 individuals and kidney = 27 individuals), with the addition of two other tissue types chosen owing to their available sample size (n: colon = 34 individuals and liver = 27 individuals). Somatic mutations in documented cancer genes[71], as well as mutations within the 25 genes having the highest mutation frequency for that tissue type, were removed to mitigate the influence of driving cancer genes on the analysis. Samples with a mutation burden greater than 3 s.d. from the mean mutation burden of that tissue were discarded. We created a shared feature set for training all clock models, selecting the 100,000 CpG sites with the greatest average somatic mutation burden across samples within ±10 kb of the CpG site.

Mutation clock

The mutation clocks were based on a gradient boosted tree model, an XGBoost Regressor[72] with default parameters, which we trained to predict chronological age using features derived from somatic mutations at the 100,000 CpG sites described above. A separate model was trained for the tumor (n = 1,601) and normal (n = 111) samples because of the different rates of mutation accumulation and mutational patterns, as described in previous literature[3,73]. The features used to describe an individual sample were (1) the counts of all somatic mutations within ±10 kb of each CpG site ('mutation burden'; Fig. 4a), (2) the one-hot-encoded tissue type, (3) the genome-wide mutation

burden, summing the MAFs across all 100,000 CpG sites, and (4) the mutation burden associated with each possible trinucleotide context (for example, CCT > CTT or TCT > TTT). To assess the influence of the whole-genome features (3 and 4), we also constructed models with only local mutation counts and tissue type (Extended Data Fig. 7). When the genome-wide features were removed from the models, the performance of the methylation clocks was unchanged (tumor: Pearson's $r = 0.83$; normal: Pearson's $r = 0.92$; Extended Data Fig. 7a,b); the mutation clocks remained significantly predictive of age (tumor: Pearson's $r = 0.53$, $P = 1.54 \times 10^{-113}$; normal: Pearson's $r = 0.73$, $P = 9.8 \times 10^{-08}$; Extended Data Fig. 7a,b); and the two measures continued to agree in the over-/underprediction of age in the same individuals (tumor: partial correlation = 0.51, $P = 6.66 \times 10^{-105}$; normal: partial correlation = 0.52, $P = 5.32 \times 10^{-04}$; Extended Data Fig. 7c,d). For the tumor tissues, the accuracy of age prediction was assessed using nested cross-validation, in which 64% of samples were used for model training, 16% for hyperparameter tuning and 20% for testing, with this entire procedure repeated over five folds (Fig. 4b and Extended Data Figs. 5 and 7). As relatively few normal tissue samples were available ($n = 111$), we assessed the performance of the normal tissue mutation clock only on the normal samples for which we had both mutation and methylation profiles ($n = 40$) to maximize the number of samples used in training while not reducing the number with calculable mutation and methylation ages. Here, we conducted a similar nested cross-validation procedure in which, in each fold, we held out 20% of the normal samples with both data types for testing ($n = 8$ samples each fold) and performed training and hyperparameter tuning on the remaining samples ($n = 71$ samples with only mutation data and $n = 32$ samples with both data types). Following hyperparameter tuning, the number of features selected for use in the trained tumor mutation clock model ranged from 1,189 to 1,293 with a mean of 1,237.2 across folds and from 307 to 344 with a mean of 322 across folds for the normal tissue mutation clock.

Methylation clock

The methylation clock was also based on an XGBoost Regressor model, with identical parameters[72] as for the mutation clock described above. A separate model was trained for the tumor ($n = 1,601$) and normal ($n = 187$) samples. The definition of features was also closely matched, focusing on the same 100,000 CpG sites but using methylation rather than mutation data. In particular, the features used to describe an individual sample were (1) the methylation fraction of each CpG, (2) the one-hot-encoded tissue type and (3) the overall degree of genome-wide methylation, computed as the sum of methylation fraction values across all 100,000 CpG sites (Fig. 4a,c and Extended Data Fig. 5). Models without the global feature were also constructed for comparison (Extended Data Fig. 7). Nested cross-validation was performed as for the mutation clock. Following hyperparameter tuning within each fold of cross-validation, the number of features selected for use in the trained tumor methylation clock ranged from 5,131 to 5,221 with a mean of 5,171.8 across cross-validation folds and from 1,987 to 2,139 with a mean of 2,051 for the normal methylation clock.

Application of existing clocks

The Hannum, Horvath and PhenoAge clocks were refitted to the TCGA pan-cancer dataset. This step was necessary as some CpG sites included in these clocks did not have methylation values passing quality controls (see 'Data access and preprocessing' in the Methods). The remaining 66%, 63% and 61% of features were used in the refitted Hannum, Horvath and PhenoAge clocks, respectively. Briefly, the CpGs used in these clocks were obtained from the respective publications[17,18,21], and an elastic-net regression model[74] with default parameters was trained on these features to predict the chronological age of TCGA samples. Nested fivefold cross-validation was used to assess the performance of each of the previously published clocks before refitting (Extended Data Fig. 6a) and all clocks (including the clock trained in this study) after being refitted to the TCGA data (Extended Data Fig. 6b). For each sample, the residual of each methylation clock was compared to the residual of the mutation clock (Fig. 4e).

Local association of methylation, mutation burden and age

Across the 1,601 individuals and 100,000 CpG loci used in the tumor mutation and methylation clocks, we compared the association between the methylation fraction of each CpG site, its mutation burden in the surrounding 20 kb and the chronological age of the samples (Fig. 4f). First, we calculated the Pearson's correlation coefficient between chronological age and the quantile-normalized methylation fraction and mutation burden (±10 kb) at each CpG locus. Second, we selected the 1%, 5% and 10% of CpG sites with the largest methylation–age correlation and mutation burden–age correlation and counted the number of CpGs shared between these groups. Third, we compared this overlap to the expected rate of overlap assuming random selection from the 100,000 original CpGs. A two-sided binomial test was applied to assess statistical significance.

Statistics and reproducibility

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 sizes. 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) greater than 3 s.d. 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.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data analyzed were from The Cancer Genome Atlas Pan-Can cohort[34–36] (http://xena.ucsc.edu/) and the Pan-Cancer Analysis of Whole Genomes[48] (https://xenabrowser.net/datapages/?hub=https://pcawg.xenahubs.net:443). Data can be accessed from the provided links and are described further in the respective publications (https://doi.org/10.1038/ng.2764, https://doi.org/10.1038/s41586-020-1969-6)[35,37]. Data to replicate the figures in this manuscript can be found on figshare ('Somatic mutation as an explanation for epigenetic aging (Koch et al. 2024)', https://figshare.com/projects/Somatic_mutation_as_an_explanation_for_epigenetic_aging_Koch_et_al_2024_/224232)[75]. 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).

Code availability

All custom algorithms and analysis code are in the GitHub repository at https://github.com/zanekoch/MutationsAndMethylationAging/.

References

1. Szilard, L. On the nature of the aging process. *Proc. Natl Acad. Sci. USA* **45**, 30–45 (1959).

2. Cagan, A. et al. Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517–524 (2022).
3. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
4. Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature* **597**, 381–386 (2021).
5. Jaiswal, S. & Ebert, B. L. Clonal hematopoiesis in human aging and disease. *Science* **366**, eaan4673 (2019).
6. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
7. Bae, T. et al. Analysis of somatic mutations in 131 human brains reveals aging-associated hypermutability. *Science* **377**, 511–517 (2022).
8. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
9. Blagosklonny, M. V. DNA- and telomere-damage does not limit lifespan: evidence from rapamycin. *Aging (Albany NY)* **13**, 3167–3175 (2021).
10. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The hallmarks of aging. *Cell* **153**, 1194–1217 (2013).
11. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23–38 (2013).
12. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362–365 (1993).
13. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).
14. Ehrlich, M. et al. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. *Nucleic Acids Res.* **10**, 2709–2721 (1982).
15. Jabbari, K. & Bernardi, G. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene* **333**, 143–149 (2004).
16. Meaney, M. J. & Szyf, M. Environmental programming of stress responses through DNA methylation: life at the interface between a dynamic environment and a fixed genome. *Dialogues Clin. Neurosci.* **7**, 103–123 (2005).
17. Hannum, G. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
18. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
19. McCrory, C. et al. GrimAge outperforms other epigenetic clocks in the prediction of age-related clinical phenotypes and all-cause mortality. *J. Gerontol. A Biol. Sci. Med. Sci.* **76**, 741–749 (2021).
20. Lu, A. T. et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)* **11**, 303–327 (2019).
21. Levine, M. E. et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)* **10**, 573–591 (2018).
22. Li, A., Koch, Z. & Ideker, T. Epigenetic aging: biological age prediction and informing a mechanistic theory of aging. *J. Intern. Med.* **292**, 733–744 (2022).
23. Yang, J.-H. et al. Loss of epigenetic information as a cause of mammalian aging. *Cell* **186**, 305–326 (2023).
24. de Magalhães, J. P. Ageing as a software design flaw. *Genome Biol.* **24**, 51 (2023).
25. López-León, M. & Goya, R. G. The emerging view of aging as a reversible epigenetic process. *Gerontology* **63**, 426–431 (2017).
26. Ito, S. et al. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**, 1300–1303 (2011).
27. Wang, M. et al. Identification of DNA motifs that regulate DNA methylation. *Nucleic Acids Res.* **47**, 6753–6768 (2019).
28. Nachun, D. et al. Clonal hematopoiesis associated with epigenetic aging and clinical outcomes. *Aging Cell* **20**, e13366 (2021).
29. Gibbs, J. R. et al. Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.* **6**, e1000952 (2010).
30. McCartney, D. L. et al. Genome-wide association studies identify 137 genetic loci for DNA methylation biomarkers of aging. *Genome Biol.* **22**, 194 (2021).
31. Youk, J., An, Y., Park, S., Lee, J.-K. & Ju, Y. S. The genome-wide landscape of C:G > T:A polymorphism at the CpG contexts in the human population. *BMC Genomics* **21**, 270 (2020).
32. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).
33. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**, 560–561 (1980).
34. Ellrott, K. et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* **6**, 271–281 (2018).
35. Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
36. Liu, J. et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* **173**, 400–416 (2018).
37. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
38. Bibikova, M. et al. High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288–295 (2011).
39. Liu, X. et al. Metallothionein 2A (MT2A) controls cell proliferation and liver metastasis by controlling the MST1/LATS2/YAP1 signaling pathway in colorectal cancer. *Cancer Cell Int.* **22**, 205 (2022).
40. Si, M. & Lang, J. The roles of metallothioneins in carcinogenesis. *J. Hematol. Oncol.* **11**, 107 (2018).
41. Fu, J. et al. Metallothionein 1G functions as a tumor suppressor in thyroid cancer through modulating the PI3K/Akt signaling pathway. *BMC Cancer* **13**, 462 (2013).
42. Tong, M. et al. Evaluation of MT family isoforms as potential biomarker for predicting progression and prognosis in gastric cancer. *Biomed Res. Int.* **2019**, 2957821 (2019).
43. Pinney, S. E. Mammalian non-CpG methylation: stem cells and beyond. *Biology (Basel)* **3**, 739–751 (2014).
44. Mathelier, A. et al. Cis-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biol.* **16**, 84 (2015).
45. Luo, X. et al. Effects of DNA methylation on TFs in human embryonic stem cells. *Front. Genet.* **12**, 639461 (2021).
46. Wang, M., Ngo, V. & Wang, W. Deciphering the genetic code of DNA methylation. *Brief. Bioinform.* **22**, bbaa424 (2021).
47. Villicaña, S. & Bell, J. T. Genetic impacts on DNA methylation: research findings and future perspectives. *Genome Biol.* **22**, 127 (2021).
48. Russo, G. et al. DNA damage and repair modify DNA methylation and chromatin domain of the targeted locus: mechanism of allele methylation polymorphism. *Sci. Rep.* **6**, 33222 (2016).
49. Morano, A. et al. Targeted DNA methylation by homology-directed repair in mammalian cells. Transcription reshapes methylation on the repaired gene. *Nucleic Acids Res.* **42**, 804–821 (2014).
50. Allen, B., Pezone, A., Porcellini, A., Muller, M. T. & Masternak, M. M. Non-homologous end joining induced alterations in DNA methylation: a source of permanent epigenetic change. *Oncotarget* **8**, 40359–40372 (2017).
51. Pagès-Gallego, M. et al. Direct detection of 8-oxo-dG using nanopore sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/2024.05.17.594638 (2024).

52. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663–676 (2006).
53. Gill, D. et al. Multi-omic rejuvenation of human cells by maturation phase transient reprogramming. *eLife* **11**, e71624 (2022).
54. Ocampo, A. et al. In vivo amelioration of age-associated hallmarks by partial reprogramming. *Cell* **167**, 1719–1733 (2016).
55. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
56. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
57. Li, R. et al. A body map of somatic mutagenesis in morphologically normal human tissues. *Nature* **597**, 398–403 (2021).
58. Chen, Y. et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).
59. Jaffe, A. E. & Irizarry, R. A. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* **15**, R31 (2014).
60. Tomusiak, A. et al. Development of an epigenetic clock resistant to changes in immune cell composition. *Commun. Biol.* **7**, 934 (2024).
61. Wang, T. et al. Quantitative translation of dog-to-human aging by conserved remodeling of the DNA methylome. *Cell Syst.* **11**, 176–185 (2020).
62. Lu, A. T. et al. Universal DNA methylation age across mammalian tissues. *Nat. Aging* **3**, 1144–1166 (2023).
63. Rozenblit, M. et al. Evidence of accelerated epigenetic aging of breast tissues in patients with breast cancer is driven by CpGs associated with polycomb-related genes. *Clin. Epigenetics* **14**, 30 (2022).
64. Moqri, M. et al. PRC2-AgeIndex as a universal biomarker of aging and rejuvenation. *Nat. Commun.* **15**, 5956 (2024).
65. Van Egeren, D. et al. Reconstructing the lineage histories and differentiation trajectories of individual cancer cells in myeloproliferative neoplasms. *Cell Stem Cell* **28**, 514–523 (2021).
66. Ferrall-Fairbanks, M. C. et al. Progenitor hierarchy of chronic myelomonocytic leukemia identifies inflammatory monocytic-biased trajectory linked to worse outcomes. *Blood Cancer Discov.* **3**, 536–553 (2022).
67. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
68. Nassar, L. R. et al. The UCSC Genome Browser database: 2023 update. *Nucleic Acids Res.* **51**, D1188–D1195 (2023).
69. Raney, B. J. et al. Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**, 1003–1005 (2014).
70. Kent, W. J. et al. The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
71. Tang, G., Cho, M. & Wang, X. OncoDB: an interactive online database for analysis of gene expression and viral infection in cancer. *Nucleic Acids Res.* **50**, D1334–D1339 (2022).
72. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, 2016).
73. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
74. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
75. Koch, Z. Zip of all data. *figshare* https://doi.org/10.6084/m9.figshare.27270468.v1 (2024).

Acknowledgements

This study was funded by the National Institutes of Health under awards U54 CA274502 (T.I.), P41 GM103504 (T.I.) and R01AG059416 (S.C.). S.C. and D.E. also receive support from The Sequoia Center for Research on Aging, California Pacific Medical Center Research Institute.

Author contributions

Z.K. designed the study, carried out the primary data analyses and wrote the manuscript. A.L. and D.S.E. assisted with data analysis and study design considerations. T.I. and S.C. designed the study and wrote the manuscript.

Competing interests

T.I. is a cofounder of Serinus and Data4Cure, is on their scientific advisory boards and has an equity interest in both companies. T.I. is on the scientific advisory board of IDEAYA Biosciences and has an equity interest. The terms of these arrangements have been reviewed and approved by the University of California, San Diego, in accordance with its conflict of interest policies. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43587-024-00794-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43587-024-00794-x.

Correspondence and requests for materials should be addressed to Steven Cummings or Trey Ideker.

Peer review information *Nature Aging* thanks Wolfgang Wagner and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in Nature Aging 2025. Koch, Zane; Li, Adam; Evans, Daniel S.; Cummings, Steven; Ideker, Trey. The dissertation author was the primary investigator and author of this paper.

Chapter 2 CHARACTERIZE PAN-EPIGENETIC ALTERATIONS ACROSS SPECIES

Abstract

Epigenetic remodeling is a hallmark of aging, yet which epigenetic layers are most affected during aging – and the extent to which they are interrelated – is not well understood. Here, we perform a comprehensive analysis of epigenetic aging encompassing 6 histone marks and DNA methylation measured across 12 tissues from >1,000 humans and mice. We identify a synchronized pattern of age-related changes across these epigenetic layers, with all changes converging upon a common set of genes. Notably, an epigenetic clock based on these genes can accurately predict age using data from any layer (Spearman ρ: 0.70 in humans, 0.81 in mice). Applying this "pan-epigenetic" clock, we observe that histone modification and DNA methylation profiles agree in prediction of which individuals are aging more rapidly or slowly. These results demonstrate that epigenetic modifications are subject to coordinated remodeling over the lifespan, offering a unified view of epigenetic aging.

Introduction

Among the hallmarks of aging, considerable attention has been devoted to the study of epigenetic alterations, particularly changes in DNA methylation patterns [1]. Across the human lifespan, thousands of cytosine-guanine dinucleotides (CpGs) undergo systematic gains or losses of methylation [2–5]. These widespread shifts have enabled the development of robust methylation-based age predictors, termed "epigenetic clocks," [5–9] which have been used pervasively in clinical and forensic applications [10–12]. Remarkably, these age-related methylation changes are highly conserved across diverse tissues and species [13], such that a single epigenetic clock can predict the age of over 100 distinct mammalian species [14].

DNA methylation (DNAm) represents only one facet of the epigenome, however. Another major class of epigenetic mark relates to chemical modifications of histones [15], which regulate chromatin structure, DNA accessibility, and gene expression. The number of distinct types of histone marks is vast and includes histone methylation, acetylation, and phosphorylation at multiple sites – some of which have known functions while many others remain incompletely understood [16]. For instance, monomethylation of lysine 4 of histone 3 (H3K4me1) commonly marks enhancers poised for activation [17], whereas H3K27me3 is associated with heterochromatic domains and transcriptional repression [18]. The occupancy of particular histone marks across the genome has been extensively investigated via chromatin immunoprecipitation sequencing (ChIP-seq), which uses DNA sequencing to read out the genomic regions that co-precipitate with a particular modified histone [19]. The combined influence of histone modifications and DNA methylation ultimately determines chromatin accessibility – the physical openness of chromatin to transcription factors and regulatory proteins [20].

Recent studies suggest that age-related changes occur not only in CpG methylation [21–23], but in some of these other epigenetic layers including histone modifications [24–27] and chromatin accessibility [28,29]. Such multifaceted change aligns with the concept that the various epigenetic processes form a tightly interconnected regulatory network, whereby alterations in one layer could reverberate through others [30]. Outside of the aging field, general studies of histone marks and DNA methylation have shown that these layers exert reciprocal effects on one another [30], driven by direct physical interactions and the shared regulatory activity of DNA methyltransferases and histone-modifying enzymes [31,32]. In studies of aging, however, each type of epigenetic mark has been typically analyzed in isolation, without relation to other epigenetic layers.

Here, we explore to what extent the lifetime changes seen in DNA methylation and histone modifications are interrelated. By analyzing epigenetic profiles from hundreds of humans and mice, we observe that the different epigenetic layers show coordinated changes during aging, with changes in all layers converging on a common set of genes. We find that an epigenetic clock based on these genes can predict age using data from any epigenetic layer, with predictions from each layer agreeing in which individuals are aging more slowly or more rapidly.

Results

Epigenetic layers show coordinated changes during aging

To characterize epigenetic change during aging, we collected histone modification (ChIP-seq) and DNA methylation (whole-genome bisulfite sequencing; WGBS) profiles from nine previous studies [33–41]. These 3,491 profiles were drawn from 482 humans and 523 mice with each individual donor contributing profiles from one or more of seven epigenetic layers across up to twelve tissues (average 3.5 profiles per donor; **Supplementary Fig. 1**). Human donors ranged from cord blood and fetal samples to older adults (0 – 85 years old), while mouse donors spanned from approximately one week of age to 35 months, collectively covering a broad range of developmental and aging time points in both species. Each histone profile reflected the proportion of cells bearing that particular histone modification at each genomic locus. For DNAm, each profile encoded the percent of cells in the measured tissue for which each CpG was methylated. To harmonize features across studies, all measured sites within the same gene body were pooled and their signal averaged to produce a single continuous value per epigenetic layer per gene (**Fig. 1a, Methods**).

Correlating the epigenetic signal in each gene with age, we found that certain epigenetic layers exhibited more age-related change than others. For H3K27me3, we observed a significant association of the occupancy of this epigenetic mark with age for 86.3% of human genes, whereas for H3K9me3, a significant age-association was seen for only 0.1% of genes (**Supplementary Fig. 2a-b**). Nevertheless, there was a clear coordination between the age-related changes among epigenetic layers, with repressive marks H3K27me3, H3K9me3, and DNA methylation aligned in their direction of change, and activating marks (H3K36me3, H3K4me1) exhibiting an opposing pattern (**Fig. 1b-c**). Unexpectedly, while H3K27ac and H3K4me3 had been largely associated with the activation of transcription, in our aging study they showed a modest but significant trend in the same direction as the repressive marks (**Fig. 1c**). Overall, we found that for 14 of the 21 possible pairs of layers, the direction of age-related change in one layer anticipated that of another (**Fig. 1d-i, Supplementary Fig. 3a-d**). For instance, DNA hypomethylation tended to arise at genes that simultaneously lost H3K9me3, whereas genes accruing H3K9me3 tended to become hypermethylated (**Fig. 1d**). In contrast, H3K4me1 and DNAm exhibited the opposite relationship, with an increased deposition of one associating with a loss of the other (**Fig. 1e**). Collectively, these age-related changes shifted each epigenetic layer towards a more disordered, higher-entropy state, as has been observed previously for DNA methylation[5] (**Supplementary Fig. 2c**).

Changes in all epigenetic layers converge upon a common set of genes

We found that the genes exhibiting the strongest associations with age were largely consistent across epigenetic layers (**Fig. 2a**). For example, the genes for which DNA methylation state was most correlated with age were highly likely to be the same genes for which H3K9me3

occupancy was correlated with age (29-fold enrichment, $p = 2 \times 10^{-98}$; **Fig. 2a**). This same convergence among epigenetic layers was seen in mice (**Fig. 2b**) involving largely the same genes as in humans (**Fig. 2c**). In both species, the strongest coordination was seen for H3K4me3 and H3K27ac, although nearly all epigenetic layers showed significant convergence on a common set of genes. At a false discovery rate (FDR) of 20%, a total of 88 genes exhibited age-related changes across all seven epigenetic layers (**Fig. 2d**).

Genes exhibiting the greatest epigenetic repression with age – those that accumulated repressive marks and lost activating marks – were highly enriched for developmental functions related to HOX gene expression and organ formation (**Fig. 2e, Methods**). Such genes included *SOX1*, *WNT10B*, and *HOXB5*. In contrast, genes with the greatest epigenetic activation with age were related to MAP kinase signalling pathways (**Fig. 2f**), including *MAP2K2* and *MTURN*.

Taken together, these analyses (**Figure 1, 2**) indicated that aging is accompanied by a coordinated shift in histone modifications and DNA methylation, with both activating and repressive marks changing together at shared genomic loci. While the direction of change varied by epigenetic mark, the loci that underwent the largest modifications during aging were conserved across species, suggestive of an intertwined "pan-epigenetic" process that unfolds over the lifespan.

Comparison of epigenetic layers in quantitative "clock" prediction of age

Having identified core genes at which aging reflects coordinated changes across epigenetic layers, we next asked whether these various layers are equally predictive of age. To address this question, we trained a collection of "single-layer" clocks, each designed to estimate age using genes as input features, with values of each gene drawn from one particular epigenetic layer

(**Fig. 3a**). Of the seven layers, H3K36me3 and H3K9me3 were omitted from clock-related analyses due to sparse coverage in mice, yielding a total of five epigenetic clock models (**Supplementary Fig. 1c**). Using a cross-validation procedure to assess these models (**Methods**), we found that the prediction accuracy varied by epigenetic layer (lowest ρ: H3K4me1 = 0.62; highest ρ: DNAm = 0.91, **Supplementary Fig. 4a-c**). Training single-mark clocks for increasing numbers of donors revealed that some marks had a greater age-prediction capacity than others, even when the number of training donors was matched (**Supplementary Fig. 4d**). As the training sample size increased, H3K27me3 and DNAm showed the most rapid improvements in model performance (scaling rate: H3K27me3 = 0.123, DNAm = 0.122; **Supplementary Fig. 4d**, **Methods**), while H3K4me1 showed the slowest (H3K4me1 scaling rate = 0.058). On the other hand, each of these single-layer clocks exhibited similar performance whether applied to humans or mice (**Supplementary Fig. 4a-c**), suggesting that, like DNA methylation [13,14], histone modifications follow an evolutionary conserved trajectory of change during aging.

We next compared the results of the single-layer clocks to a "pan-epigenetic" clock designed to predict donor age from any of the five epigenetic layers analyzed (**Fig. 3a**, **Methods**). This model had an architecture identical to the single-mark clocks but was trained using profiles from all epigenetic layers. This pan-epigenetic clock accurately predicted age in both humans (Spearman ρ = 0.70; **Fig. 3b**) and mice (Spearman ρ = 0.81, **Fig. 3c**) from any epigenetic mark (**Fig. 3d**). Notably, the performance of the pan-epigenetic clock closely matched that of the separate single-layer clocks in each species (Spearman ρ = 0.90, $p = 3.4 \times 10^{-04}$; **Fig. 3e**) and tissue (Spearman ρ = 0.98, $p = 3.1 \times 10^{-08}$; **Fig. 3f**). When we randomized all profiles from one epigenetic layer during training, the model largely retained the ability to estimate the age of non-randomized, held-out donors using that particular layer (**Fig. 3g**, **Methods**). These results demonstrated that the

aging signals contained within each epigenetic layer are similar enough to be effectively represented by a single model, with each layer reinforcing predictions from others.

Different epigenetic layers exhibit synchronized aging rates

If there is a coordinated process driving epigenetic change across layers, individuals appearing older/younger than their chronological ages according to one layer would likewise appear older/younger according to other layers. To test this hypothesis, we applied the pan-epigenetic clock to donors profiled for all five epigenetic modifications (n = 109 human donors). We found that the degree of over- or under-prediction of chronological age was indeed synchronized across layers (**Fig. 4a-d**, **Methods**). For instance, for individuals whose H3K27ac profile indicated that they were 1 year older than their chronological age, their H3K4me3 profile produced an overprediction of 0.5 ± 0.1 years (mean ± s.d., **Fig. 4a**). This synchronization of the rate of epigenetic change between modifications was not limited to associations within histone marks, with DNAm patterns also yielding age predictions that were significantly associated with those of histone marks (**Fig. 4c,d**).

Reduced methylation age is strongly promoted by chronic caloric restriction [42] (CR), which has long been known to extend lifespan in many species [43]. We thus evaluated whether this intervention would likewise slow the progression of age-related changes across the various layers of the epigenome. Applying the pan-epigenetic clock to tissues from C57BL/6 mice fed a CR or control diet [33,35], we found that every layer tended to produce younger age predictions in CR mice, but that these reductions reached statistical significance for only DNAm and H3K4me1 (**Fig. 4e-f**). Ages inferred from DNAm and H3K4me1 profiles of CR mice were on

average 4.1 and 0.7 months younger than their chronological ages, respectively. Thus, CR appears to exert a youth-preserving influence on multiple layers of the epigenome.

Discussion

Collectively, our results support three general findings regarding the links between the multiple epigenetic marks and age. First, epigenetic layers in both humans and mice follow a shared trajectory of age-related change – converging on the same genomic loci and advancing at synchronized rates (**Figures 1, 2**). Second, this coherence allows a single pan-epigenetic clock to predict chronological age from any layer in either species (**Figure 3**). Third, within each individual, the epigenetic age inferred from one mark is echoed by similar age predictions based on every other mark (**Figure 4**).

Recent evidence points to extensive interaction among epigenetic regulatory enzymes [44]. Such crosstalk provides a potential molecular explanation underlying the coordination of age-related changes we observed across epigenetic layers. For instance, DNA methyltransferase 1 (*DNMT1*) selectively methylates loci marked by H3K9me3 [45] – aligning with our finding of age-associated DNAm gains at regions with increasing H3K9me3 (**Fig. 1d, g**). Similarly, physical interaction between the histone acetyltransferase p300 and the SET1 family of methyltransferases promotes cooperative H3K4me3 and H3K27ac deposition [46], which is likewise reflected in our results (**Fig. 1e, h**). Notably, the positive correlation between H3K27ac and DNA methylation during aging is counter-intuitive, given that H3K27ac is canonically associated with transcriptional activation while DNA methylation at promoters is generally repressive; one possible interpretation is that H3K27ac does not directly promote DNA methylation, but rather that both marks are independently recruited to the same genomic loci as part of a coordinated,

multi-layer epigenetic remodeling program during aging, potentially driven by shared upstream regulators or chromatin states. Age-related changes impacting one epigenetic layer may reverberate through the entire epigenetic network, advancing epigenetic age across all layers. Conversely, a therapeutic intervention that rejuvenates a single epigenetic layer could possibly restore youthful epigenetic states throughout multiple layers of the epigenome, similar to the effect we observed in calorically restricted mice (**Fig. 4e-f**).

As to what are the initial causes of the observed pan-epigenetic alterations, two main theories have emerged. The first, a developmental or "pseudo-programmatic" model [47], posits that aging represents a continuation of the epigenetic trajectories established during development. This view is supported by the remarkable conservation of age-associated epigenetic shifts across species and the enrichment of these changes for developmental genes [13,14,48]. These patterns are supported by the present results, in which repressive epigenetic marks converge on genes governing development, in both mice and humans (**Supplementary Fig. 4a**). The second theory – the stochastic drift model – suggests that aging-associated epigenetic alterations arise from the random accumulation of molecular errors and damage-induced changes throughout the epigenome. Supporting this theory, computational simulations have demonstrated that much of the epigenetic change seen during aging can be explained by the random accumulation of epigenetic alterations [49,50], which may result from DNA damaging events [51,52]. Our observation of increasing entropy across all layers of the epigenome during aging (**Supplementary Fig. 2c**) dovetails with this second model as well. The two different theories are not strictly incompatible [53], such that epigenetic aging may arise from both inherited developmental programs and the accumulation of molecular noise, with the epigenetic crosstalk reported here amplifying alterations from either source (**Fig. 1**).

Some limitations of this study are as follows. First, our results are based on cross-sectional data: future studies applying longitudinal profiling will be essential to disentangle true age-related changes from cohort effects. Second, while this study has expanded the palette of epigenetic marks associated with aging, it is not exhaustive in this respect – inclusion of additional modifications and histone variants could refine the observed synchrony. Third, despite stringent batch correction (**Methods**), differences in the ChIP-seq and WGBS protocols among the studies from which we drew data may affect our results. Fourth, relative-age scaling relies on imperfect maximum lifespan estimates which may limit comparability of aging trajectories across species. Fifth, because the cohort spans a wide age range that includes infant and juvenile donors, the epigenetic changes captured by our models likely reflect a combination of developmental and aging processes rather than aging alone.

Online Methods

Histone modification processing. ChIP–seq narrowPeak files were gathered from six public datasets – ENCODE, BLUEPRINT, CEEHRC, Signal et al. 2024, Yang et al. 2023, and Hillje et al. 2022. The locations of autosomal gene bodies in humans and mice were retrieved from the Ensembl v109 catalogue [54]. Then, for each histone profile separately, the mean narrowPeak score of all peaks overlapping a given gene was taken as the epigenetic value of that gene, while genes without any overlapping peaks in that sample were assigned a value of zero. Mouse gene identifiers were converted to their one-to-one human orthologs via *pybiomart [55]*, placing every profile in a common human-gene coordinate system. Genes with no overlapping peaks in any sample were discarded. This procedure yielded gene-level histone modification profiles aligned between species.

DNA methylation processing. Whole-genome bisulfite sequencing files were obtained from six public datasets – ENCODE, BLUEPRINT, CEEHRC, Stubbs et al. 2017, Meer et al. 2018, and Petkovich et al. 2017. CpG sites were intersected with autosomal gene bodies defined in Ensembl v109 via *pybedtools* [56]. For each sample, the mean methylation fraction of all CpGs falling within a given gene was recorded. Genes lacking overlapping CpGs were assigned a value of zero. Mouse gene identifiers were mapped to their one-to-one human orthologs using *pybiomart* [55], placing every methylome on the same human-gene coordinate system.

Integration and normalization of epigenetic layers. Gene-level histone mark and DNA methylation datasets were first scaled separately. Within each {dataset × epigenetic-mark} stratum: (1) epigenetic signal values were min–max normalised to the interval [0, 1] and (2) , missing values (<0.5% of all values) were imputed as the mean across all other samples. Next, all datasets, including both histone marks and DNA methylation, were combined. Genes for which epigenetic values were present in all datasets were retained (n = 17,602). To mitigate inter-study batch effects, the combined matrix was processed with the empirical-Bayes method from pyComBat[57,58], using "dataset" as the batch factor and leaving biological covariates unmodelled. The resulting batch-corrected matrix served as the input for all downstream analyses.

Calculating epigenome entropy. Gene-level epigenetic signal intensities from the batch-corrected matrix were used to compute Shannon entropy. For each sample, absolute signal intensities across the 17,602 genes were normalized to sum to one, yielding a discrete probability distribution. Shannon entropy was then calculated as:

$$H = \sum_{i=0}^{i=17,602} p_i \times log_2(p_i)$$

In this calculation p_i denotes the normalized value of gene i; genes with zero value were omitted from the calculation.

Overlap of age-associated regions across epigenetic layers and species. Within the batch-corrected matrix, gene-wise associations between chronological age and epigenetic values were quantified separately for every {species × epigenetic-mark} stratum using Spearman correlation. Ranking the correlation values by their unsigned magnitude, pairwise overlaps among the 1,000 genes with the strongest age association was computed for all histone-mark combinations within each species (**Fig. 1d-e**). The expected overlap under random expectation was taken to be:

$$Expected\ overlap\ =\ N_{top} \times \left[\frac{N_{top}}{G}\right]^2$$

Where N_{top} = 1,000 genes and G = 17,602 (the total number of genes). Fold enrichment was calculated as the observed divided by expected overlap. Statistical significance was assessed with two-sided binomial tests followed by Bonferroni correction. Concordance between human and mouse age-associated genes was evaluated analogously (**Fig. 1f**).

Interaction between age-related changes among epigenetic layers. For each pair of epigenetic layers *A* and *B*: (1) the 1,000 genes with the strongest positive (age-increasing set) or negative (age-decreasing set) age-association (Spearman ρ) in layer A were selected; (2) the distribution of Spearman ρ's in layer *B* was extracted for the genes in the age-increasing and age-decreasing sets of layer *A*; (3) the median difference in Spearman ρ's between these sets (age-increasing – age-decreasing) was calculated; and (4) a two-sided Mann–Whitney U test was used to assess if there existed statistical significance of medians between the distribution of Spearman ρ values in the age-increasing vs. age decreasing sets. These p values were Bonferroni-corrected across all (7 choose 2 =) 21 comparisons (α = 0.05). A significant median difference signified cross-talk between epigenetic layers, with positive values indicating that genes gaining signal in layer A with age also tend to gain signal in layer B, whereas negative values denoted opposing trends (**Fig. 1g-i, Supplementary Fig. 3a-d**).

Enrichment analyses. Genes showing the strongest age-associated epigenetic repression were defined by taking the intersection of the top 1,000 most positively age-associated genes according to each repressive mark (H3K27me3, H3K9me3, DNAm) and the top 1,000 most most negatively age-associated genes according to each activating mark (H3K27ac, H3K36me3, H3K4me1, H3K4me3) in each species (n = 778 genes). The converse was done to define the most epigenetically activated genes with age (n = 427 genes). Enrichments of these genes for Gene Ontology Biological Process and Reactome pathways were measured using the Enrichr API via the gseapy [59] package, testing against the "GO_Biological_Process_2023" and "Reactome_Pathways_2024" libraries, respectively. The universe for each test comprised all genes quantified in the age–correlation analysis. For each gene set (repressed or activated), terms containing fewer than two genes or with Benjamini–Hochberg–adjusted p-value ≥ 0.05 were excluded. The Combined Score from Enrichr (log p × z-score) was used as the enrichment metric.

Representation of chronological age. Developmental and aging time points were aligned between humans and mice by representing the age of each donor as a percentage of the maximum lifespan of its respective species, as described by Lu et al. 2023 [14]. Briefly, each donor's age (in years) was divided by the species-specific maximum lifespan reported in GenAge [60] (120 years for humans; 4 years for mice) to obtain a unitless relative age between 0 and 1. These values were then transformed with the negative log–log transformation:

$$Scaled\ age\ = -log(-log(\frac{chronological\ age}{maximum\ lifespan})$$

The reverse transformation was used to present values in units of years in all figures.

Clock dataset. Epigenetic profiles originating from cell lines, sex-specific tissues (prostate, breast or other reproductive organs), placenta and any tissue represented by < 50 libraries were excluded from both model training and evaluation. Calorically restricted samples (n = 298 mice)

were likewise withheld from model fitting but retained for subsequent inference. Due to their limited representation in mice (**Supplementary Fig. 1c**), H3K36me3 and H3K9me3 profiles were omitted from all clock construction steps, leaving five epigenetic layers (H3K4me3, H3K27ac, H3K27me3, H3K4me1, and DNA-methylation). The resulting dataset comprised 2,598 profiles from 1,005 unique donors.

Single-layer clock training. Five independent gradient boosted decision-tree regressors [61] were fit – one for each epigenetic layer (H3K4me3, H3K27ac, H3K27me3, H3K4me1 and DNA-methylation) – to predict the age of mouse and human samples. Models used the 17,602 genes retained after layer integration as predictors (see "Integration and normalization of epigenetic layers"); no additional covariates were included. Training employed stratified ten-fold cross-validation in which the number of profiles from donors of each age quintile was balanced across folds, while ensuring that all profiles from the same donor resided in a single fold. Default XGBoost hyper-parameters were used, and performance was computed over the five held-out folds.

Pan-epigenetic clock training. The pan-epigenetic clock was likewise an XGBoost model, trained in an identical way to the single-layer clocks. The same feature set and samples were used. However, instead of samples from each epigenetic layer being trained on and predicted by separate clocks, a single model was trained over all profiles.

Permutation analysis of cross-layer learning. To determine whether the pan-epigenetic clock leverages shared aging information across layers, we retrained the model five times, each time randomly permuting one epigenetic layer while leaving all others intact. For every sample belonging to the layer under permutation, we independently randomized the gene-level signal values among genes within that sample; this preserves the distribution of intensities but destroys

the correspondence between each gene and its true value. The scrambled matrix was used only in the training folds, whereas the held-out folds retained the original, non-scrambled values, for prediction. Using an identical model and cross-validation process, a clock was fit to each permuted dataset and then its performance was assessed based on its prediction of the non-permuted, held-out samples.

Scaling rate of age-prediction. Using donors from the CEEHRC, ENCODE, and BLUEPRINT datasets who had been profiled for multiple epigenetic marks (n = 219 donors), single-layer clocks were trained on training cohorts of varying size. For each mark, we repeatedly (ten times) sampled $20-200$ donors in 20-donor increments without replacement, trained a single-layer clock on each training set exactly as described above, and recorded the age-prediction accuracy (Spearman ρ). The scaling rate for a given epigenetic layer was taken as the slope of a zero-intercept ordinary-least-squares regression of these Spearman ρ values on the $\log_{10}$-transformed number of training donors. The log transformation was used as the scaling rate was nonlinear.

Software. All analyses were performed in Python 3.10. 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.

Data availability

Data can be accessed through the respective publications: Canadian Epigenetics, Environment and Health Research Consortium (CEEHRC) [41], ENCODE [40], Blueprint [39], Signal et al. 2024 [38], Yang et al. 2023 [37], Stubbs et al. 2017 [36], Hillje et al. 2022 [35], Meer et al. 2018 [34], Petkovich et al. 2017 [33].

Code availability

All custom algorithms and analysis code is in the GitHub repository at https://github.com/zanekoch/pan-epigenetic-age-prediction.

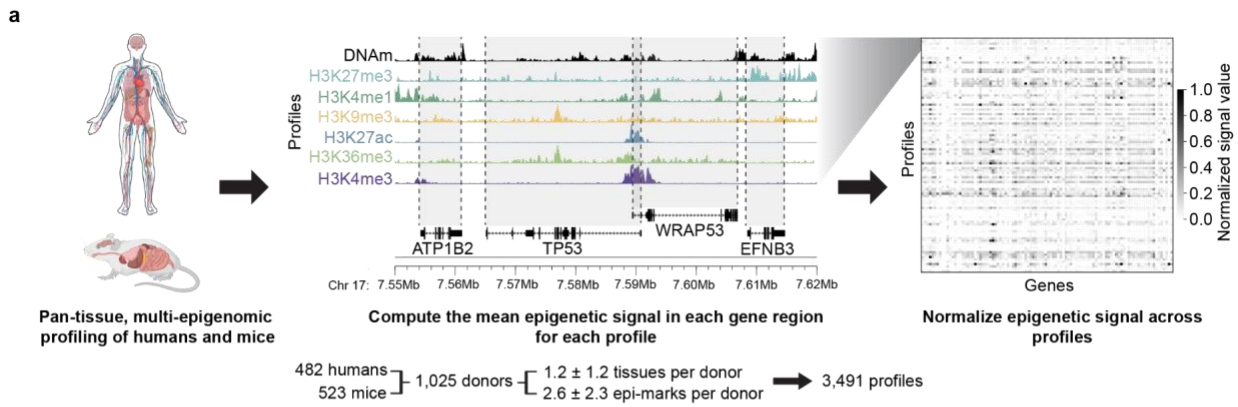
Author Contributions

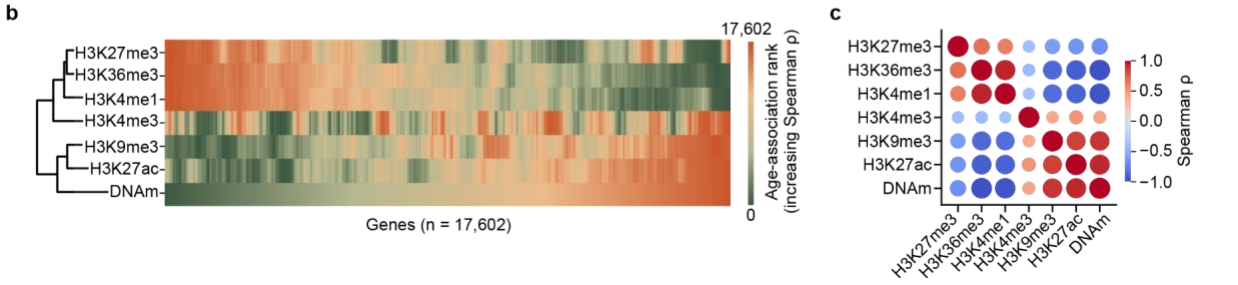
ZK and AL designed the study, carried out the primary data analyses, and wrote the manuscript. TI designed the study and wrote the manuscript.

Figures

Figure 5

a) Workflow for generating gene-level epigenetic features. The mean signal value within each gene region for profiles from human and mouse donors was calculated, batch corrected, and normalized across profiles to create a feature matrix.
b) Heatmap of the association (Spearman's ρ) between the signal of each epigenetic layer and age across 17,602 genes (n = 482 human donors). Genes (columns) are ordered by increasing DNA methylation age-association and colored by the rank of association within each layer, with green denoting the most negative correlations and orange the most positive.
c) Heatmap of pairwise Spearman correlations (ρ) between gene-level age associations for each epigenetic mark; all correlations were statistically significant ($p < 1\times10^{-10}$, two-sided Spearman correlation test). Circle size and color representing the strength and direction of correlations.
d) Kernel density estimate (KDE) of Spearman's ρ between DNA methylation and age for all genes (grey) versus the 1,000 genes with the most positive (light color) or most negative (dark color) H3K9me3 age associations. P-value was calculated based on a two-sided Mann–Whitney U (MWU) test comparing the distribution DNAm age-association values between the sets of positively and negatively H3K9me3 age-associated genes.
e) Similar to (d), but stratifying H3K4me1 age-associations based on DNAm levels.
f) Similar to (d), but stratifying H3K4me3 age-associations based on H3K27ac levels.
g-i), For each focal mark – DNAm (g), H3K4me1 (h), and H3K4me3 (i) – genes were ranked by their Spearman age-association (ρ) according to each of the other epigenetic layers (x-axis) and split into the 1,000 most positively and 1,000 most negatively associated. Bars show the median difference in ρ of the focal mark between these two gene sets. P values (two-sided Mann–Whitney U test) are shown above each bar; "n.s." denotes $p \geq 0.05$.


a
DNAm
H3K27me3
H3K4me1
H3K9me3
H3K27ac
H3K36me3
H3K4me3
Profiles
ATP1B2
TP53
WRAP53
EFNB3
Chr 17: 7.55Mb 7.56Mb 7.57Mb 7.58Mb 7.59Mb 7.60Mb 7.61Mb 7.62Mb
Genes
Normalized signal value
Pan-tissue, multi-epigenomic profiling of humans and mice
Compute the mean epigenetic signal in each gene region for each profile
Normalize epigenetic signal across profiles
482 humans
523 mice
1,025 donors
1.2 ± 1.2 tissues per donor
2.6 ± 2.3 epi-marks per donor
3,491 profiles


b
17,602
Age-association rank (increasing Spearman ρ)
0
Genes (n = 17,602)
c
Spearman ρ

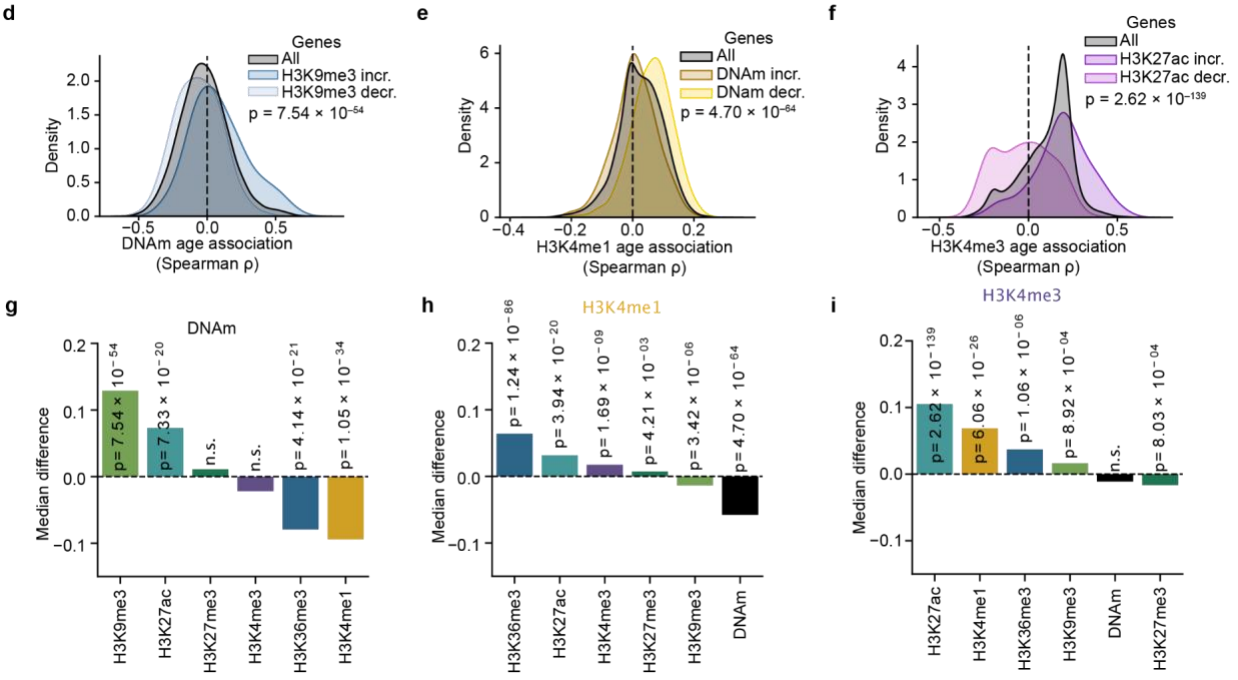

d
Genes
All
H3K9me3 incr.
H3K9me3 decr.
p = 7.54 × 10⁻⁵⁴
Density
DNAm age association (Spearman ρ)
e
All
DNAm incr.
DNam decr.
p = 4.70 × 10⁻⁶⁴
H3K4me1 age association (Spearman ρ)
f
All
H3K27ac incr.
H3K27ac decr.
p = 2.62 × 10⁻¹³⁹
H3K4me3 age association (Spearman ρ)
g
DNAm
Median difference
p = 7.54 × 10⁻⁵⁴
p = 7.33 × 10⁻²⁰
n.s.
n.s.
p = 4.14 × 10⁻²¹
p = 1.05 × 10⁻³⁴
H3K9me3
H3K27ac
H3K27me3
H3K4me3
H3K36me3
H3K4me1
h
H3K4me1
p = 1.24 × 10⁻⁸⁶
p = 3.94 × 10⁻²⁰
p = 1.69 × 10⁻⁰⁹
p = 4.21 × 10⁻⁰³
p = 3.42 × 10⁻⁰⁶
p = 4.70 × 10⁻⁶⁴
H3K36me3
H3K27ac
H3K4me3
H3K27me3
H3K9me3
DNAm
i
H3K4me3
p = 2.62 × 10⁻¹³⁹
p = 6.06 × 10⁻²⁶
p = 1.06 × 10⁻⁰⁶
p = 8.92 × 10⁻⁰⁴
n.s.
p = 8.03 × 10⁻⁰⁴
H3K27ac
H3K4me1
H3K36me3
H3K9me3
DNAm
H3K27me3

Figure 6

a) Upset plot showing the fold enrichment for the overlap between the top 1,000 most age-associated human genes (absolute Spearman's ρ) for each pair of epigenetic marks compared to random expectation (n = 482 human donors, **Methods**). (*) indicates $p < 0.005$ based on a two-sided binomial test.
b) Similar to (a), but for age-associations calculated in mice (n = 523 mouse donors).
c) Bar plot depicting the fold enrichment of the top 1,000 age-associated genes for each epigenetic mark between mice and humans. (*) indicates $p < 5.0 \times 10^{-36}$ based on a two-sided binomial test.
d) Nested circles representing the number of genes found to be significantly associated with age in all seven epigenetic layers at three different FDR thresholds (n = 482 human donors). The radius of each circle is proportional to $\log_{10}$ of the number of genes represented by that circle.
e) Gene Ontology Biological Process enrichment scores for genes showing the strongest epigenetic repression (teal bars, left of zero) versus activation (magenta bars, right of zero) with age. Processes are ordered by enrichment score on the x-axis, with increased enrichment indicated by values far from zero. Representative genes contributing to each term are listed alongside each bar.
f) Reactome pathway enrichment plotted as in (e).

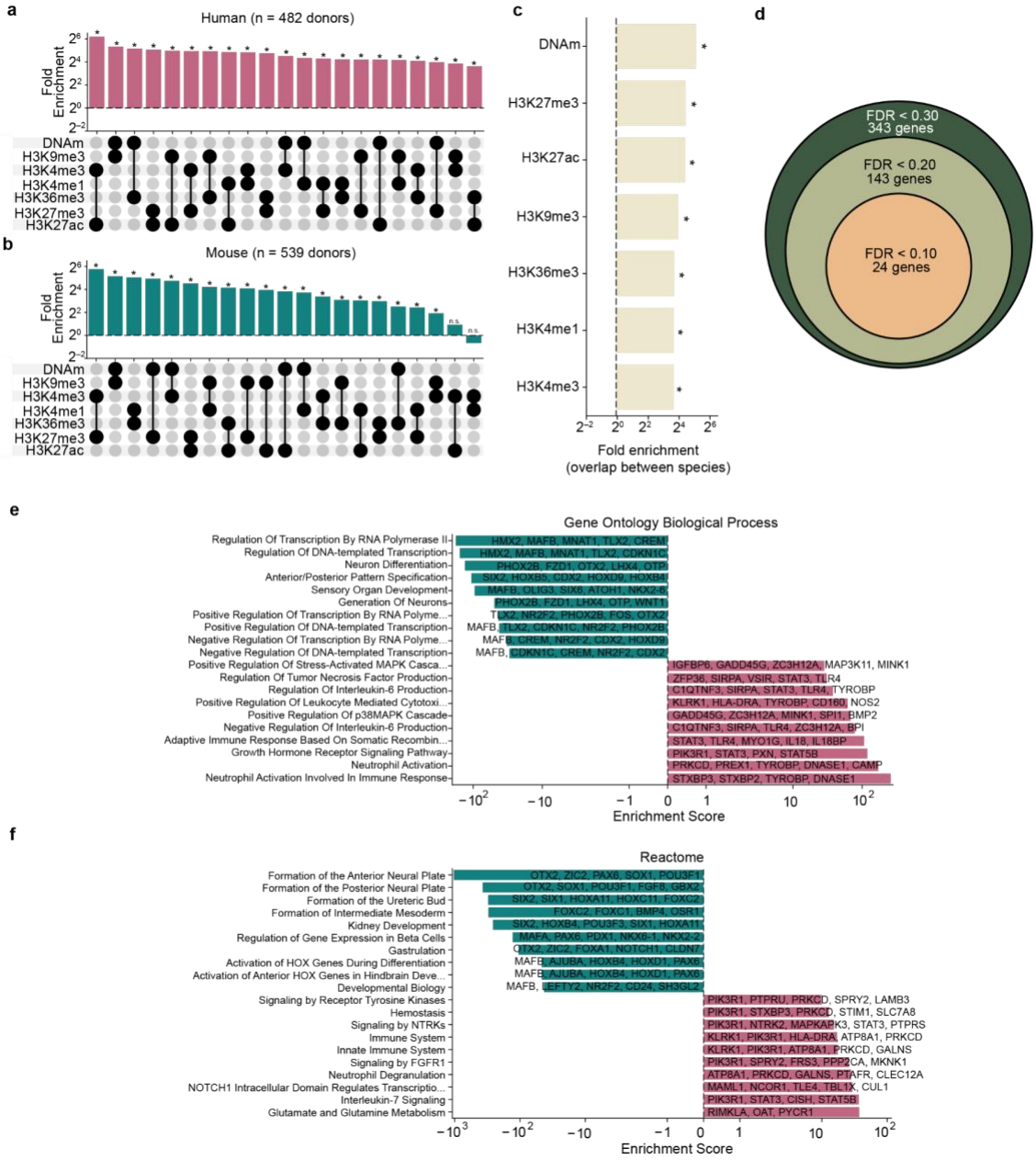

a
Human (n = 482 donors)
Fold Enrichment
DNAm
H3K9me3
H3K4me3
H3K4me1
H3K36me3
H3K27me3
H3K27ac
b
Mouse (n = 539 donors)
Fold Enrichment
n.s.
DNAm
H3K9me3
H3K4me3
H3K4me1
H3K36me3
H3K27me3
H3K27ac
c
DNAm
H3K27me3
H3K27ac
H3K9me3
H3K36me3
H3K4me1
H3K4me3
Fold enrichment
(overlap between species)
d
FDR < 0.30
343 genes
FDR < 0.20
143 genes
FDR < 0.10
24 genes
e
Gene Ontology Biological Process
Regulation Of Transcription By RNA Polymerase II
Regulation Of DNA-templated Transcription
Neuron Differentiation
Anterior/Posterior Pattern Specification
Sensory Organ Development
Generation Of Neurons
Positive Regulation Of Transcription By RNA Polyme...
Positive Regulation Of DNA-templated Transcription
Negative Regulation Of Transcription By RNA Polyme...
Negative Regulation Of DNA-templated Transcription
Positive Regulation Of Stress-Activated MAPK Casca...
Regulation Of Tumor Necrosis Factor Production
Regulation Of Interleukin-6 Production
Positive Regulation Of Leukocyte Mediated Cytotoxi...
Positive Regulation Of p38MAPK Cascade
Negative Regulation Of Interleukin-6 Production
Adaptive Immune Response Based On Somatic Recombin...
Growth Hormone Receptor Signaling Pathway
Neutrophil Activation
Neutrophil Activation Involved In Immune Response
Enrichment Score
f
Reactome
Formation of the Anterior Neural Plate
Formation of the Posterior Neural Plate
Formation of the Ureteric Bud
Formation of Intermediate Mesoderm
Kidney Development
Regulation of Gene Expression in Beta Cells
Gastrulation
Activation of HOX Genes During Differentiation
Activation of Anterior HOX Genes in Hindbrain Deve...
Developmental Biology
Signaling by Receptor Tyrosine Kinases
Hemostasis
Signaling by NTRKs
Immune System
Innate Immune System
Signaling by FGFR1
Neutrophil Degranulation
NOTCH1 Intracellular Domain Regulates Transcriptio...
Interleukin-7 Signaling
Glutamate and Glutamine Metabolism
Enrichment Score

Figure 7

a) Two complementary age-prediction strategies pursued in this study: five "single-mark" clocks trained to predict age with a single epigenetic modification (left) versus a unified "pan-epigenetic" clock trained to predict from any of the five marks (right). Donor ages are represented as a percentage of the species' maximum lifespan.
b) Scatter plot of predicted versus actual age for human donors (n = 482 donors, n = 2,029 profiles) using the pan-epigenetic clock. Each point represents an individual donor and error bars the standard deviation of age predictions across samples from each tissue profiled in that donor. Predictions based on different epigenetic marks are shown in different colors.
c) Similar to (b) but for mouse donors (n = 523 mice, n = 569 profiles).
d) Radar plot comparing the predictive performance (Spearman ρ) of the pan-epigenetic clock on samples of each epigenetic modification and species (human: orange polygon, n = 482 donors; mouse: grey polygon, n = 523 donors).
e) Scatter plot comparing the predictive performance of the single-mark clocks to that of the pan-epigenetic clock. Each point represents the Spearman correlation between predicted and actual age for a particular epigenetic mark (indicated by color, same as panel b) in a particular species (indicated by shape).
f) Spearman correlation between predicted and actual age in each tissue (human: n = 482 donors; mouse: n = 523 donors). Bars denote the Spearman ρ across all epigenetic mark types based upon the pan-epigenetic clock (solid bars) and the single-mark clocks (hatched bars).
g) Heatmap showing the effect of permutation on predictive accuracy of the pan-epigenetic clock. Each circle denotes the percent change in age-prediction performance (relative to un-permuted) using the epigenetic mark on the y-axis when the epigenetic mark on the x-axis is randomized before model training (**Methods**).

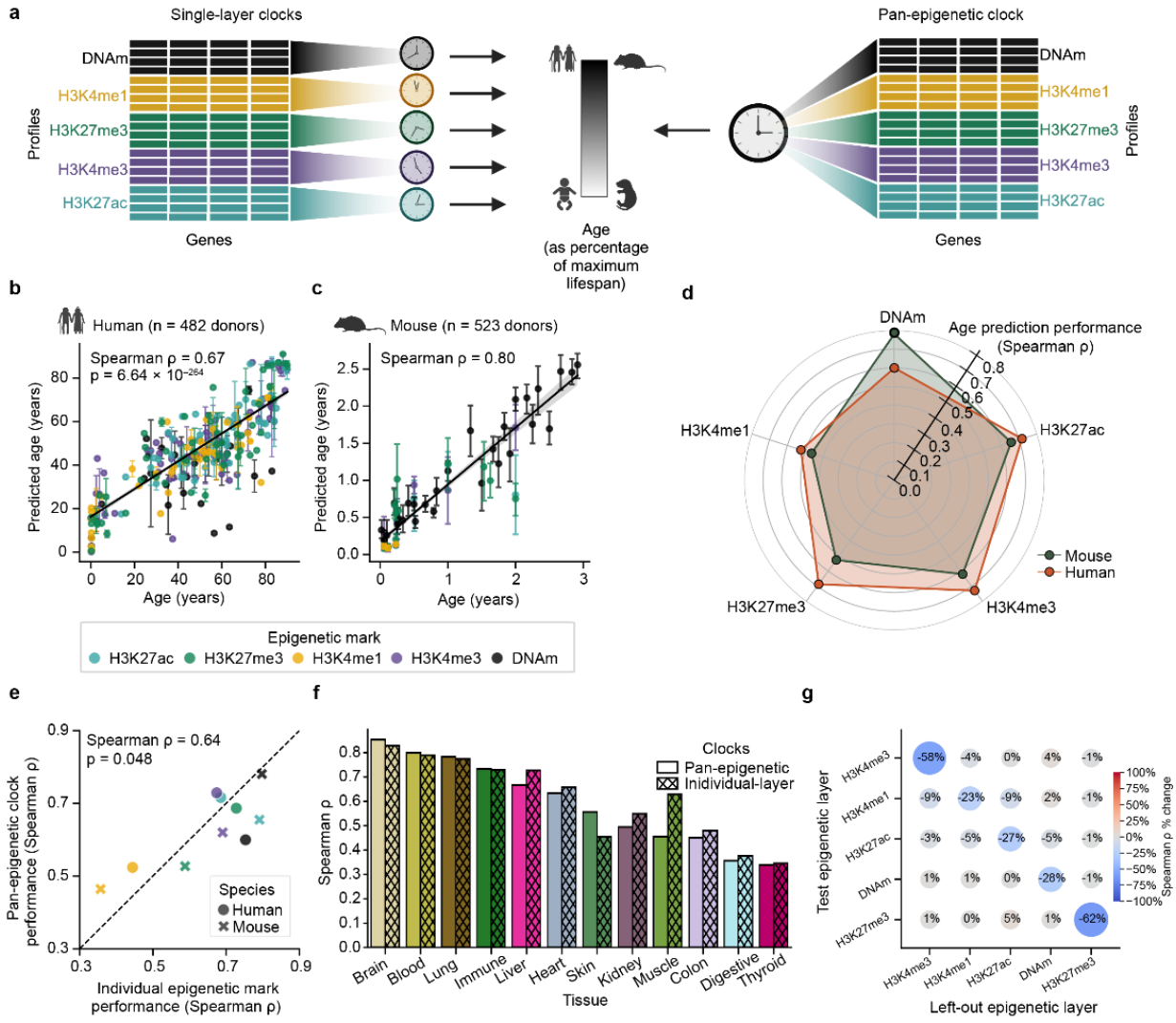
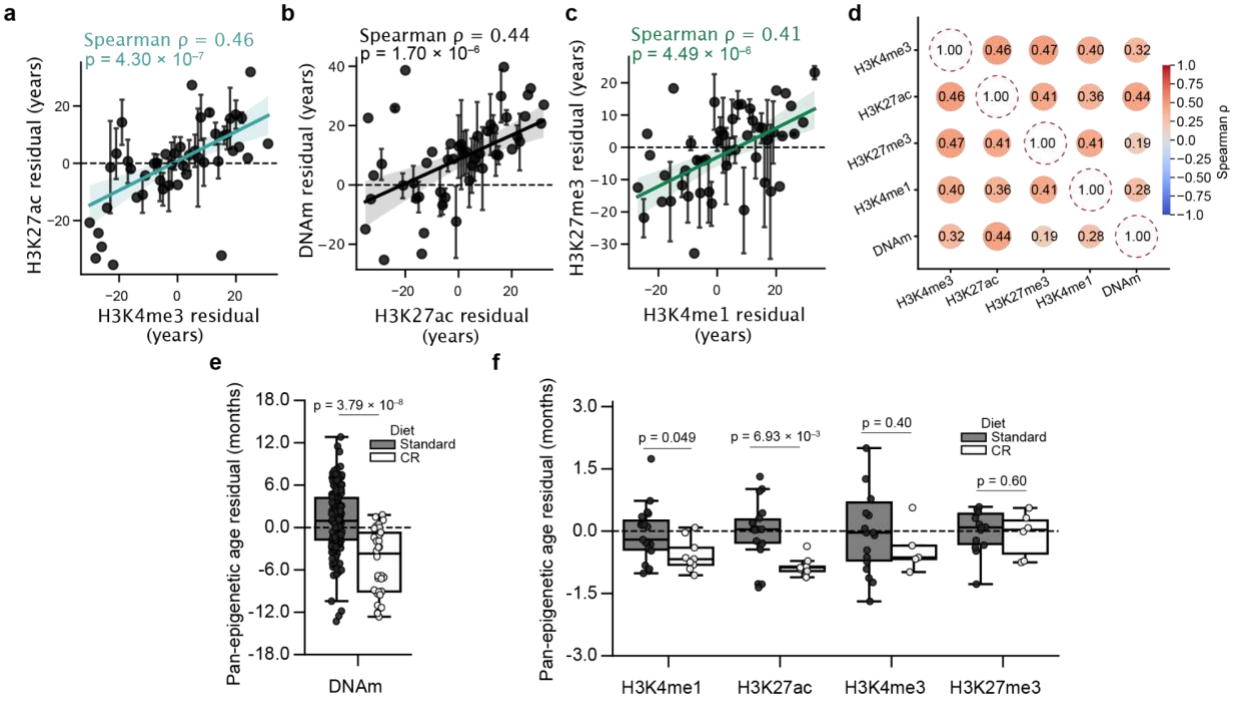

a
Single-layer clocks
DNAm
H3K4me1
H3K27me3
H3K4me3
H3K27ac
Profiles
Genes
Age
(as percentage
of maximum
lifespan)
Pan-epigenetic clock
DNAm
H3K4me1
H3K27me3
H3K4me3
H3K27ac
Profiles
Genes
b
Human (n = 482 donors)
Spearman ρ = 0.67
p = 6.64 × 10−264
Predicted age (years)
Age (years)
c
Mouse (n = 523 donors)
Spearman ρ = 0.80
Predicted age (years)
Age (years)
Epigenetic mark
H3K27ac H3K27me3 H3K4me1 H3K4me3 DNAm
d
Age prediction performance
(Spearman ρ)
DNAm
H3K4me1
H3K27ac
H3K27me3
H3K4me3
Mouse
Human
e
Spearman ρ = 0.64
p = 0.048
Pan-epigenetic clock
performance (Spearman ρ)
Individual epigenetic mark
performance (Spearman ρ)
Species
Human
Mouse
f
Clocks
Pan-epigenetic
Inidividual-layer
Spearman ρ
Brain
Blood
Lung
Immune
Liver
Heart
Skin
Kidney
Muscle
Colon
Digestive
Thyroid
Tissue
g
Test epigenetic layer
H3K4me3
H3K4me1
H3K27ac
DNAm
H3K27me3
-58% -4% 0% 4% -1%
-9% -23% -9% 2% -1%
-3% -5% -27% -5% -1%
1% 1% 0% -28% -1%
1% 0% 5% 1% -62%
Left-out epigenetic layer
Spearman ρ % change

Figure 8

a) Violin plots comparing age prediction residuals between H3K27ac and H3K4me3 (n = 109 human donors). ρ, Spearman correlation between the two residuals. P-value denotes significance calculated by modeling Spearman ρ's as a Student's t distribution. Inner box plots denote median and interquartile range.
b) Similar to (a), but comparing H3K27me3 and H3K4me1.
c) Similar to (a), but comparing DNAm and H3K4me3.
d) Correlation matrix illustrating the relationship between age prediction residuals (i.e., partial correlation; **Methods**) from the pan-epigenetic clock across five epigenetic modifications (n = 109 human donors), with circle size and color representing the strength and direction of Spearman correlations.
e) Box plots depicting the pan-epigenetic age residuals (in months) from the DNAm of mice under standard (n = 175 mice) versus caloric restriction (n = 32 mice) diets. Points indicate individual mice. P value calculated using a two-sided Mann–Whitney U test.
f) Similar to (e), but age predictions made using histone marks (n = 62 CR mice; n = 29 control fed mice).


a
Spearman ρ = 0.46
p = 4.30 × 10⁻⁷
H3K27ac residual (years)
H3K4me3 residual (years)
b
Spearman ρ = 0.44
p = 1.70 × 10⁻⁶
DNAm residual (years)
H3K27ac residual (years)
c
Spearman ρ = 0.41
p = 4.49 × 10⁻⁶
H3K27me3 residual (years)
H3K4me1 residual (years)
d
H3K4me3
H3K27ac
H3K27me3
H3K4me1
DNAm
1.00 0.46 0.47 0.40 0.32
0.46 1.00 0.41 0.36 0.44
0.47 0.41 1.00 0.41 0.19
0.40 0.36 0.41 1.00 0.28
0.32 0.44 0.19 0.28 1.00
Spearman ρ
e
p = 3.79 × 10⁻⁸
Diet
Standard
CR
Pan-epigenetic age residual (months)
DNAm
f
p = 0.049
p = 6.93 × 10⁻³
p = 0.40
p = 0.60
Diet
Standard
CR
Pan-epigenetic age residual (months)
H3K4me1
H3K27ac
H3K4me3
H3K27me3

References

1. Gonzalo S. Epigenetic alterations in aging. J Appl Physiol. 2010;109: 586–597.

2. Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML. Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci U S A. 2005;102: 10604–10609.

3. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL. Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet. 2009;5: e1000602.

4. Li A, Koch Z, Ideker T. Epigenetic aging: Biological age prediction and informing a mechanistic theory of aging. J Intern Med. 2022;292: 733–744.

5. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S. Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell. 2013;49: 359–367.

6. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14: R115.

7. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S. An epigenetic biomarker of aging for lifespan and healthspan. Aging . 2018;10: 573–591.

8. Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K. DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging . 2019;11: 303–327.

9. Belsky DW, Caspi A, Corcoran DL, Sugden K, Poulton R, Arseneault L. DunedinPACE, a DNA methylation biomarker of the pace of aging. Elife. 2022;11. doi:10.7554/eLife.73420

10. Bell CG, Lowe R, Adams PD, Baccarelli AA, Beck S, Bell JT. DNA methylation aging clocks: challenges and recommendations. Genome Biol. 2019;20: 249.

11. Higgins-Chen AT, Thrush KL, Wang Y, Minteer CJ, Kuo P-L, Wang M. A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking. Nat Aging. 2022;2: 644–661.

12. Paparazzo E, Lagani V, Geracitano S, Citrigno L, Aceto MA, Malvaso A. An ELOVL2-Based Epigenetic Clock for Forensic Age Prediction: A Systematic Review. Int J Mol Sci. 2023;24. doi:10.3390/ijms24032254

13. Wang T, Ma J, Hogan AN, Fong S, Licon K, Tsui B. Quantitative Translation of Dog-to-Human Aging by Conserved Remodeling of the DNA Methylome. Cell Syst. 2020;11: 176–185.e6.

14. Lu AT, Fei Z, Haghani A, Robeck TR, Zoller JA, Li CZ. Universal DNA methylation age across mammalian tissues. Nat Aging. 2023;3: 1144–1166.

15. Geiman TM, Robertson KD. Chromatin remodeling, histone modifications, and DNA methylation-how does it all fit together? J Cell Biochem. 2002;87: 117–125.

16. Zhao Y, Garcia BA. Comprehensive Catalog of Currently Documented Histone Modifications. Cold Spring Harbor Perspectives in Biology. 2015. p. a025064. doi:10.1101/cshperspect.a025064

17. Spicuglia S, Vanhille L. Chromatin signatures of active enhancers. Nucleus. 2012;3: 126–131.

18. Cai Y, Zhang Y, Loh YP, Tng JQ, Lim MC, Cao Z. H3K27me3-rich genomic regions can function as silencers to repress gene expression via chromatin interactions. Nat Commun. 2021;12: 719.

19. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10: 669–680.

20. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nat Rev Genet. 2019;20: 207–220.

21. Occean JR, Yang N, Sun Y, Dawkins MS, Munk R, Belair C. Gene body DNA hydroxymethylation restricts the magnitude of transcriptional changes during aging. Nat Commun. 2024;15: 6357.

22. Johnson ND, Huang L, Li R, Li Y, Yang Y, Kim HR. Age-related DNA hydroxymethylation is enriched for gene expression and immune system processes in human peripheral blood. Epigenetics. 2020;15: 294–306.

23. Chien J-F, Liu H, Wang B-A, Luo C, Bartlett A, Castanon R. Cell-type-specific effects of age and sex on human cortical neurons. Neuron. 2024;112: 2524–2539.e5.

24. de Lima Camillo LP, Asif MH, Horvath S, Larschan E, Singh R. Histone mark age of human tissues and cell types. Sci Adv. 2025;11: eadk9373.

25. Sen P, Dang W, Donahue G, Dai J, Dorsey J, Cao X. H3K36 methylation promotes longevity by enhancing transcriptional fidelity. Genes Dev. 2015;29: 1362–1376.

26. Greer EL, Maures TJ, Hauswirth AG, Green EM, Leeman DS, Maro GS. Members of the H3K4 trimethylation complex regulate lifespan in a germline-dependent manner in C. elegans. Nature. 2010;466: 383–387.

27. Maures TJ, Greer EL, Hauswirth AG, Brunet A. The H3K27 demethylase UTX-1 regulates C. elegans lifespan in a germline-independent, insulin-dependent manner. Aging Cell. 2011. pp. 980–990. doi:10.1111/j.1474-9726.2011.00738.x

28. Bozukova M, Nikopoulou C, Kleinenkuhnen N, Grbavac D, Goetsch K, Tessarz P. Aging is associated with increased chromatin accessibility and reduced polymerase pausing in liver. Mol Syst Biol. 2022;18: e11002.

29. Rechsteiner C, Morandini F, Perez K, Praz V, López-García G, Hinte L. Development of a novel aging clock based on chromatin accessibility. bioRxiv. 2022. p. 2022.08.11.502778. doi:10.1101/2022.08.11.502778

30. Fu K, Bonora G, Pellegrini M. Interactions between core histone marks and DNA methyltransferases predict DNA methylation patterns observed in human cells and tissues. Epigenetics. 2020;15: 272–282.

31. Zhang Y, Jurkowska R, Soeroes S, Rajavelu A, Dhayalan A, Bock I. Chromatin methylation activity of Dnmt3a and Dnmt3a/3L is guided by interaction of the ADD domain with the histone H3 tail. Nucleic Acids Res. 2010;38: 4246–4253.

32. Lehnertz B, Ueda Y, Derijck AAHA, Braunschweig U, Perez-Burgos L, Kubicek S. Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin. Curr Biol. 2003;13: 1192–1200.

33. Petkovich DA, Podolskiy DI, Lobanov AV, Lee S-G, Miller RA, Gladyshev VN. Using DNA Methylation Profiling to Evaluate Biological Age and Longevity Interventions. Cell Metab. 2017;25: 954–960.e6.

34. Meer MV, Podolskiy DI, Tyshkovskiy A, Gladyshev VN. A whole lifespan mouse multi-tissue DNA methylation clock. Elife. 2018;7. doi:10.7554/eLife.40675

35. Hillje R, Luzi L, Amatori S, Persico G, Casciaro F, Rusin M. Time makes histone H3 modifications drift in mouse liver. Aging (Albany NY). 2022;14: 4959–4975.

36. Stubbs TM, Bonder MJ, Stark A-K, Krueger F, BI Ageing Clock Team, von Meyenn F. Multi-tissue DNA methylation age predictor in mouse. Genome Biol. 2017;18: 68.

37. Yang N, Occean JR, Melters DP, Shi C, Wang L, Stransky S. A hyper-quiescent chromatin state formed during aging is reversed by regeneration. Mol Cell. 2023;83: 1659–1676.e11.

38. Signal B, Phipps AJ, Giles KA, Huskins SN, Mercer TR, Robinson MD. Ageing-related changes to H3K4me3, H3K27ac, and H3K27me3 in purified mouse neurons. Cells. 2024;13: 1393.

39. Fernández JM, de la Torre V, Richardson D, Royo R, Puiggròs M, Moncunill V. The BLUEPRINT Data Analysis Portal. Cell Syst. 2016;3: 491–495.e5.

40. ENCODE Project Consortium, Moore JE, Purcaro MJ, Pratt HE, Epstein CB, Shoresh N. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature. 2020;583: 699–710.

41. Bujold D, Morais DA de L, Gauthier C, Côté C, Caron M, Kwan T. The International Human Epigenome Consortium Data Portal. Cell Syst. 2016;3: 496–499.e2.

42. Wang T, Tsui B, Kreisberg JF, Robertson NA, Gross AM, Yu MK. Epigenetic aging signatures in mice livers are slowed by dwarfism, calorie restriction and rapamycin treatment. Genome Biol. 2017;18: 57.

43. Speakman JR, Mitchell SE. Caloric restriction. Mol Aspects Med. 2011;32: 159–221.

44. Lempiäinen JK, Garcia BA. Characterizing crosstalk in epigenetic signaling to understand disease physiology. Biochem J. 2023;480: 57–85.

45. Ren W, Fan H, Grimm SA, Guo Y, Kim JJ, Yin J. Direct readout of heterochromatic H3K9me3 regulates DNMT1-mediated maintenance DNA methylation. Proc Natl Acad Sci U S A. 2020;117: 18439–18447.

46. Tang Z, Chen W-Y, Shimada M, Nguyen UTT, Kim J, Sun X-J. SET1 and p300 act synergistically, through coupled histone modifications, in transcriptional activation by p53. Cell. 2013;154: 297–310.

47. Gems D, Virk RS, de Magalhães JP. Epigenetic clocks and programmatic aging. Ageing Res Rev. 2024;101: 102546.

48. Moqri M, Cipriano A, Simpson DJ, Rasouli S, Murty T, de Jong TA. PRC2-AgeIndex as a universal biomarker of aging and rejuvenation. Nat Commun. 2024;15: 5956.

49. Meyer DH, Schumacher B. Aging clocks based on accumulating stochastic variation. Nat Aging. 2024;4: 871–885.

50. Tong H, Dwaraka VB, Chen Q, Luo Q, Lasky-Su JA, Smith R. Quantifying the stochastic component of epigenetic aging. Nat Aging. 2024;4: 886–901.

51. Koch Z, Li A, Evans DS, Cummings S, Ideker T. Somatic mutation as an explanation for epigenetic aging. Nat Aging. 2025;5: 709–719.

52. Yang J-H, Hayano M, Griffin PT, Amorim JA, Bonkowski MS, Apostolides JK. Loss of epigenetic information as a cause of mammalian aging. Cell. 2023;186: 305–326.e27.

53. Tarkhov AE, Lindstrom-Vautrin T, Zhang S, Ying K, Moqri M, Zhang B. Nature of epigenetic aging from a single-cell perspective. Nat Aging. 2024;4: 854–870.

54. Harrison PW, Amode MR, Austine-Orimoloye O, Azov AG, Barba M, Barnes I. Ensembl 2024. Nucleic Acids Res. 2024;52: D891–D899.

55. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G. BioMart--biological queries made easy. BMC Genomics. 2009;10: 22.

56. Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics. 2011;27: 3423–3424.

57. Behdenna A, Colange M, Haziza J, Gema A, Appé G, Azencott C-A. pyComBat, a Python tool for batch effects correction in high-throughput molecular data using empirical Bayes methods. BMC Bioinformatics. 2023;24: 459.

58. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020;2: lqaa078.

59. Fang Z, Liu X, Peltz G. GSEApy: a comprehensive package for performing gene set enrichment analysis in Python. Bioinformatics. 2023;39: btac757.

60. de Magalhães JP, Abidi Z, Dos Santos GA, Avelar RA, Barardo D, Chatsirisupachai K. Human Ageing Genomic Resources: updates on key databases in ageing research. Nucleic Acids Res. 2024;52: D900–D908.

61. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. arXiv [cs.LG]. 2016. Available: http://arxiv.org/abs/1603.02754

Acknowledgements

Chapter 2 contains unpublished material coauthored with Koch, Zane; Li, Adam.; and Ideker, Trey. The dissertation author was the primary author of this chapter.

Chapter 3 TEST DREAM COMPLEX ACTIVITY AS A DETERMINANT OF GENOMIC AGING RATE

Abstract

The DREAM complex has emerged as a central repressor of DNA repair, raising questions as to whether such repression exerts long-term effects on human health. Here we establish that DREAM activity significantly impacts lifetime somatic mutation burden, and that such effects are linked to altered lifespan and age-related disease pathology. First, joint profiling of DREAM activity and somatic mutations across a single-cell atlas of 21 mouse tissues shows that cellular niches with lower DREAM activity have decreased mutation rates. Second, DREAM activity predicts the varied lifespans observed across 92 mammals, with low activity marking longer-lived species. Third, reduced DREAM activity in Alzheimer's patients predicts late disease onset and decreased risk for severe neuropathology. Finally, we show DREAM knockout protects against mutation accumulation *in vivo*, reducing single-base substitutions by 4.2% and insertion/deletions by 19.6% in brains of mice. These findings position DREAM as a key regulator of aging.

Introduction

To protect against the constant barrage of DNA lesions arising from endogenous cellular processes and exogenous damaging agents, organisms have evolved an extensive network of DNA repair mechanisms [1]. These mechanisms are especially effective in long-lived species, which exhibit increased resistance to cellular stresses [2,3], decreased somatic mutation rates [4,5], and more efficient DNA repair [6]. Elevated DNA repair activity is also present in reproductive (germline) tissues, yielding mutation rates that are up to 100-fold lower than in the soma [7–9].

Genomic studies in centenarians have found strong enrichments for specific genetic variants that activate DNA repair genes [10–12]; conversely, genomic instability is a common feature of nearly all progeroid diseases [13,14] and has been reported to engender numerous phenotypes of old age including neuroinflammation [15], type 2 diabetes [16], kidney dysfunction [17,18], and Alzheimer's disease (AD) [19–22]. This prioritization of DNA repair in long-lived species, reproductive tissues, and centenarians underscores the importance of genomic integrity to longevity and raises questions about how such differences in DNA repair arise.

Recently, a protein complex known as DREAM has been shown to transcriptionally repress DNA repair pathways in somatic cells, where these pathways are not necessary for reproductive success [23]. The DREAM complex (named for its subcomponents, Dp, Rb-like-1, E2f, And MuvB) was originally characterized as a repressor of the cell cycle [24]. It forms as cells enter quiescence (G0 phase), when MuvB is phosphorylated by the dual-specificity tyrosine-regulated kinase 1A (DYRK1A) [25], leading to DREAM formation and the repression of numerous target genes [26,27]. To later exit quiescence, the components of the DREAM complex are prompted to dissociate by the phosphorylation of p130 (*RBL2*) by cyclin-dependent kinases [24,28], lifting repression of DREAM targets [26]. Following these cell-cycle studies, the direct targets of the DREAM complex have been shown to include many genes functioning in DNA damage response (DDR) [23,29]. Disruption of DREAM in nematode worms, progeroid mice, or human cells produces resistance to DNA damaging agents [23], whereas cells with an extra copy of DYRK1A – and thus greater DREAM activity [30] – show increased levels of γH2AX, a marker of DNA damage [31].

Collectively, these recent results have extended the role of DREAM from repression of cell cycle to repression of DDR and its associated functions. Here we seek to determine if the DREAM

complex not only governs the cell cycle and DNA repair but also hallmarks of aging and longevity (**Fig. 1a**). We find that, at the cellular level, DREAM activity associates with somatic mutation rate; across species, DREAM activity inversely correlates with maximum lifespan; in humans, diminished DREAM activity predicts later and less severe Alzheimer's disease pathology; and in mice, induced DREAM deficiency slows the rate of somatic mutation.

Results

Transcriptional readout of DREAM repressor activity

We selected genes with regulatory regions previously shown to be bound by multiple components of the DREAM complex through protein-DNA interactions [23,24] (n = 328 genes). The expression of these DREAM-target genes was used to score DREAM complex activity (**Methods**), signed such that low expression corresponds to high DREAM activity (**Fig. 1a-b**).

The utility of this DREAM activity score was underscored by several observations. First, in human tissues this score was tightly associated with the protein abundance and modification status of the DREAM-subunits MuvB and p107/p130, whose phosphorylation status controls DREAM assembly [28] (Pearson $r = 0.70$, $p = 6.8 \times 10^{-15}$, n = 160 human individuals, **Supplementary Fig. 1a-b**, **Methods**). Second, the score successfully distinguished cells treated with DREAM complex inhibitors [23], harmine or INDY, from those of control cells (**Fig. 1c**). Third, gene expression analysis of nematode worms with a mutation in *lin-52*, which causes defects in DREAM complex assembly, yielded an activity score that was markedly lower than that of wild-type worms [23] (**Fig. 1d**). Fourth, increased DREAM activity was associated with lower expression of the proliferative marker MKI67 [32] across cell types (Spearman $\rho = -0.56$, $p = 4.37 \times 10^{-11}$, n = 120 cell types, **Supplementary Fig. 1c**) and tissues (Spearman $\rho = -0.45$, $p = 0.03$, n = 23 tissues, **Supplementary Fig. 1d**) – as would be expected given the DREAM complex's role

as a negative regulator of the cell cycle [24,26]. Collectively, these results suggested that the DREAM activity score is an informative and specific metric for quantifying the repressive activity of the DREAM complex.

DREAM activity associates with somatic mutation rate in single cells

We used this DREAM activity measure to study the connection between DREAM and somatic mutation. For this purpose we analyzed the *Tabula Muris Senis* dataset (TMS, **Supplementary Table 1**) [33], which includes full-length RNA sequencing [34] of single cells collected from 21 organs in 18 C57BL/6JN mice at young (3 months), middle (18 months), and old (24 months) age. DREAM activity was scored in the single-cell RNA-seq data of each of 110,824 cells, revealing a broad range of values that varied significantly across organs and individual mice, and decreased with age (**Supplementary Fig. 2a-c**).

Somatic mutations were called from the same scRNA-seq data, leveraging a previous analysis [33] (**Methods**). Here we observed that the overall burden of somatic mutations increased with age in nearly every tissue (**Fig. 2a**), corresponding to rates of mutation ranging from a low of 0.0025 mutations/kb/month (tongue) to a high of 0.0045 (kidney).

Notably, in the vast majority of tissues, cells with high DREAM activity had significantly greater mutation burdens than those exhibiting low activity (**Fig. 2b-e**, **Supplementary Fig. 2d-e**). For example, kidney cells in the top quintile of DREAM activity, showed a 2.4-fold increase in somatic mutations per kb compared to cells in the bottom quintile ($p = 3.3 \times 10^{-49}$, n = 180 cells per quintile, two-sided Mann-Whitney U test). This positive association between DREAM activity and mutation burden was observed not only at the tissue level but also within individual cell types within a tissue. For instance, liver hepatocytes with greater DREAM activity had significantly

greater mutation burdens at each age (**Fig. 2f**), and this was the case for the majority of cell types (**Fig. 2g**) and tissues (**Fig. 2h**) even when correcting for cell proliferation rate (**Supplementary Fig. 2f-g**). In contrast to DREAM target genes, the expression levels of randomly chosen gene sets were not significantly associated with mutation burden (cell types: $p = 2.6 \times 10^{-32}$, tissue types: $p = 4.86 \times 10^{-14}$; **Fig. 2g-h**, **Methods**).

DREAM activity predicts lifespan and mutation rate across 92 species

To relate these results to lifespan, we scored DREAM activity in the transcriptomes of 92 mammalian species, broadly spanning six taxonomic orders [35] (~3 individuals × 3 tissues per species, n = 803 samples, **Methods**, **Fig. 3a**). Comparing DREAM activity scores to the maximum lifespan of each species, we observed a strong inverse relationship (**Fig. 3b**). Negative associations were observed across all tissues profiled (**Fig. 3c**) and became even stronger after controlling for the adult weight (AW) of each species (Spearman ρ: liver = –0.59, $p = 6.23 \times 10^{-24}$; brain = –0.46, $p = 1.26 \times 10^{-13}$; kidney = –0.30, $p = 6.23 \times 10^{-06}$; **Fig. 3d**, **Supplementary Fig. 3a-b**). Furthermore, DREAM activity levels tended to be coordinated across all tissues within a species (**Fig. 3e**).

Comparing the DREAM activity of liver, brain, and kidney tissues to previous measurements of species-specific somatic mutation rates [4] (n = 11 species), we found that DREAM complex activity in the liver and brain was positively correlated with somatic mutation rate across species (Liver: $\rho = 0.61$, $p = 6.0 \times 10^{-4}$; Brain: $\rho = 0.49$, $p = 0.008$; Kidney: $\rho = –0.31$, $p = 0.12$; **Fig. 3f, Supplementary Fig. 3c-g, Methods**). Further linking DREAM complex activity to species-specific DNA repair capacity, we observed that fibroblasts from species with lower DREAM complex activity exhibited greater resistance to various stress-inducing agents including

the mitochondrial respiration inhibitor rotenone; the oxidative damage inducing agents cadmium and paraquat; the alkylating agent methyl methanesulfonate (MMS); UV radiation; and heat (n = 37 cell lines from 6 species [2], **Supplementary Fig. 4a-h, Methods**). These results indicated that the amount of DREAM complex activity within a species is associated with the lifespan of that species, as well as its somatic mutation rate and resistance to DNA damage.

Low DREAM activity protects survival on a high-fat diet

A high-fat diet has been shown to shorten lifespan and induce genomic instability in multiple species [36–38], associated with an increased burden of oxidative damage and rate of somatic mutation [39]. Using a large cohort of mice from 50 distinct inbred strains [40] (n = 882 mice, **Supplementary Fig. 5a**), half fed a high-fat diet (HFD; 60% calories from fat) and half fed a regular chow diet (CD; 6% calories from fat), we investigated whether strains with lower DREAM activity were protected from the harms of HFD leading to extended survival.

For each strain and diet, 70% of mice (n = 612 mice) were followed to measure lifespan, while the remaining 30% were sacrificed and transcriptionally profiled to score DREAM activity (n = 270 mouse livers). Lifespan and DREAM activity varied greatly across strains (median lifespan: 704 ± 147 days, mean ± std; **Methods**), with HFD mice having markedly shorter lifespans than CD mice (median survival: HFD = 640 days, CD = 730 days; **Supplementary Fig. 5b-f**). Notably, strains with lower DREAM activity had extended survival on the HFD compared to those with higher DREAM activity (median survival: lowest 20% DREAM activity = 678 days, highest 20% DREAM activity = 578 days; **Fig. 4a-b**). Controlling for sequencing depth and physiological factors – including weight, weight gain on diet, and blood lipid levels (**Supplementary Fig. 5g**, **Methods**) – a one standard deviation increase in DREAM activity was associated with a 36%

increase in risk of death for HFD mice (Hazard ratio = 1.36, $p = 9.0 \times 10^{-4}$; **Fig. 4b**). When fed the CD, there was no significant difference in survival associated with DREAM activity (Hazard ratio = 0.93, p = 0.27; **Fig. 4b**, **Supplementary Fig. 5h**).

We considered whether the association between DREAM activity and lifespan in the HFD mice could be confounded by effects of the high-fat diet itself, whereby the diet increased DREAM activity while decreasing lifespan. However, this interpretation was contradicted by the observation that the DREAM complex activity of a strain when fed the CD was indicative of the survival of mice of the same strain when fed the HFD (Spearman $\rho = -0.55$, p = 0.002; **Fig. 4c**); that is, higher CD DREAM activity predicted shortened survival when challenged by the HFD. Collectively, these findings suggested that inherent differences in DREAM activity determine resilience to HFD-induced stress, with lower DREAM activity conferring a survival advantage under metabolic challenge.

DREAM activity is linked to neuropathy and Alzheimer's disease

We next moved from mice to human samples to test the relevance of our results to diseases associated with human aging. For this purpose, we applied the SComatic [41] method to identify somatic single nucleotide substitution mutations in 803,122 single cells profiled using single-cell RNA sequencing by the Tabula Sapiens [42] and Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD [43]) consortia (**Fig. 5a**, **Methods**). Cells had been drawn from 11 tissues of 90 human individuals, including 80 with Alzheimer's Disease and 10 without disease, with donors ranging from 33 to 102 years of age (85.2 ± 13.5 years, mean ± std). Somatic mutations accumulated in tissues at varying rates: from approximately 5 per cell per year in bladder and lung cells, to 10 in brain cells (**Fig. 5b**), and 22 in fat cells (**Fig. 5c, Supplementary Fig. 6a-f**). Notably, we observed

that human cells with the highest DREAM activity in each tissue obtained somatic mutations at substantially greater rates than the cells with the lowest DREAM activity (**Fig. 5d**), replicating the results we found in mice (**Fig. 2**).

Neuropathological data had been collected at the time of death for Alzheimer's diseased individuals (n = 80). We found that increased Braak score (a measure of tau pathology), Alzheimer's Disease Neuropathologic Change score (ADNC, a composite measure of amyloid plaques, neurofibrillary tangles, and neuritic plaques), and somatic mutation burden were all associated with greater DREAM activity (**Fig. 5e, Methods**). In an independent dataset of post-mortem brains from individuals with neurodegeneration (Religious Orders Study/Memory and Aging Project, ROSMAP [44], n = 1,101 individuals), elevated DREAM activity was associated with earlier AD diagnosis (**Fig. 5f**). Controlling for covariates, a one standard-deviation increase in DREAM activity corresponded to a 21% increase in the risk of AD at a given age (HR = 1.21, p = 0.01, **Fig. 5g**). Unlike DREAM-target genes, the expression of randomly selected sets of genes did not exhibit a significant association with AD diagnosis risk (mean HR = 1.0, **Supplementary Fig. 6h-i**). We also considered that reverse causation was possible: the longer an individual had AD before death, the lower their DREAM activity would become. However, DREAM activity was not associated with the duration of AD before death (**Supplementary Fig. 6j**).

These results indicated that DREAM activity is associated with human dysfunction at multiple biological scales: somatic mutations at the genomic level, neuropathology at the cellular level, and Alzheimer's disease at the organismal level.

DREAM loss of function reduces mutation accumulation *in vivo*

Finally, we tested whether DREAM activity is a causal driver of mutation accumulation *in vivo* by preventing DREAM complex assembly in mice and quantifying lifetime somatic mutation burden. We analyzed formalin-fixed (FFPE) brains from DREAM loss-of-function (LoF) [45] mice (n = 4, $p107^{D/D}$ $p130^{-/-}$) and littermate controls (n = 5; $p107^{D/D}$ $p130^{fl/fl}$, **Fig. 6a**). At 8 weeks of age, both groups had received tamoxifen, preventing DREAM assembly only in the DREAM LoF mice, while littermate controls retained intact DREAM function [45]. Brain tissue was collected after natural death and profiled for somatic mutations using single-molecule duplex sequencing (UDSeq, **Methods**).

Controlling for covariates, we found that DREAM LoF mice exhibited a 4.2% reduction in single-base substitution (SBS) mutations per cell compared to controls (**Fig. 6b**, **Methods**). Across samples, the operative mutational signatures – distinctive mutation patterns that reflect the underlying source of somatic mutations [46] – were SBS1, SBS5, SBS30, as well as the characteristic mutational signature of formalin fixation [47]. SBS1 reflects spontaneous deamination of 5-methylcytosine and SBS5 is a ubiquitous clock-like process of uncertain origin – both are classical age-related signatures – whereas SBS30 is linked to base-excision repair defects [48]. The proportion of mutations attributed to each mutational signature was not significantly different between groups, suggesting that loss of DREAM reduces overall mutagenesis rather than selectively altering a single repair pathway (**Fig. 6c-d**).

We observed that DREAM loss-of-function also conferred protection from insertion and deletion mutations (ID). Relative to controls, DREAM LoF mice had 19.6% fewer ID mutations (**Fig. 6e-f**). The ID mutational spectra also shifted: the prevalence of the ID21 mutational signature

– 2-4 bp deletions in double-repeats – was significantly reduced in DREAM LoF brains, whereas ID23 – 1 bp deletions at homopolymer runs – was increased (**Fig. 6g**). Together, these results indicated that loss of DREAM activity reduces the lifetime accrual of somatic mutations, spanning multiple mutation types and mutational processes.

Discussion

Collectively, our results suggest that the DREAM complex plays a pivotal role in modulating the rate of DNA damage accumulation, thereby shaping organismal lifespan and the onset of aging-related pathologies (**Fig. 6h**). Consistent with a causal role, genetic disruption of DREAM assembly protects mice from somatic mutation, directly linking DREAM activity to mutation accumulation in vivo (**Fig. 6**). These results corroborate, and may help explain, previous research demonstrating that long-lived species and centenarians possess mechanisms enhancing genomic stability [6,10–12].

The simplest interpretation of our findings is that when the DREAM complex represses DNA repair, DNA damage tends to remain unresolved leading to somatic mutation and other consequences. Some of our experimental data indeed support this model: for instance, loss of DREAM reduced both SBS and ID mutation rates and reshaped the ID mutational spectrum (**Fig. 6**, reducing signature ID21 [49]) consistent with enhanced repair of replication-slippage intermediates. Furthermore, DREAM target genes include those involved in nearly every major DNA repair pathway [23,24], including MUTYH, NEIL3, and POLD1 (base excision repair [50,51]); RAD51, XRCC2, and BRCA2 (homologous recombination [52]); XRCC4, PARG, and MRE11 (non-homologous end joining [53]); and POLQ, PCNA, and USP1 (translesion synthesis [54]). Simultaneously, unrepaired lesions can give rise to mutations by multiple mechanisms – such as cytosine deamination in the case of UV-induced cyclobutane pyrimidine dimers [55], mispairing of oxidized bases during DNA replication [56,57], and error-prone translesion synthesis at replication forks stalled by DNA damage [58]. Taken together, these observations suggest that by reducing the expression of DNA repair factors, the DREAM complex allows DNA damage to

persist, thereby increasing the likelihood that unrepaired lesions will be converted into permanent mutations through replication or error-prone repair processes.

Long-lived species exhibit a range of adaptations thought to underlie their extended lifespans, including enhanced DNA repair capacities [2], reduced somatic mutation rates [4,5], altered metabolic rates [59], and larger body sizes [60]. We observed an inverse association between DREAM complex activity and species lifespan (**Fig. 3**), which may be either a cause or result of such adaptations. By permitting increased expression of DNA repair genes, reduced DREAM activity in long-lived species could plausibly enhance DNA repair and decrease somatic mutation rates. This idea aligns with reports of elevated expression of DNA repair genes in long-lived species [61,62], although structural modifications to the repair factors themselves have also been implicated in such heightened repair capabilities [6]. Alternatively, reduced DREAM activity may be a downstream effect of other adaptations that more directly confer extended longevity. Although the observed relationship between DREAM activity and lifespan does not appear to be confounded by species size (**Fig. 3d**, **Supplementary Fig. 3a-b**), the possibility of other confounding factors, such as lower metabolic rate – or additional, yet unidentified mechanisms – warrants further investigation.

Within species, we found that elevated DREAM activity was associated with both the decreased survival of mice fed a high-fat diet (**Fig. 4**) and earlier onset of AD in humans (**Fig. 5**), a pattern potentially explained by heightened genomic damage. Indeed, high DREAM activity correlated with greater somatic mutation burdens and more severe neuropathology, aligning with prior evidence implicating DNA damage in age-related diseases [63–65] in general and in Alzheimer's disease [66,67] in particular. Nevertheless, we cannot exclude the possibility that in

these cases DREAM activity is merely a marker of other processes that drive morbidity and mortality, rather than a direct cause.

Extrapolating from these findings, DREAM inhibition could represent a new therapeutic approach to prevent or mitigate DNA damage-driven conditions, including aging. Compounds which disrupt DREAM complex formation [23,68,69] have already shown beneficial effects in model organisms – reducing oxidative stress [70], preserving bone density [71], enhancing insulin sensitivity [72,73], and improving cognitive function [74]. Our results suggest that the upregulation of DNA repair brought about by these DREAM inhibiting compounds may underlie their therapeutic benefits. While some of these compounds are in early development for human neurological disorders [75,76], the link between DREAM activity and DNA repair explored here points to a broader opportunity to use DREAM-modulating therapies to slow or prevent age-related pathologies in humans.

Methods

scRNA-seq processing. Each scRNA-seq dataset was processed separately using Scanpy [77]. First, any cell with fewer than 200 detected genes or more than 10,000,000 total read counts was removed. Second, any gene detected in fewer than three cells was discarded (using the Scanpy functions `filter_cells` and `filter_genes`). Third, the total read count within each cell was normalized to 10,000 (`normalize_per_cell`), log-scaled (`log1p`), and then scaled to have a value between zero and ten (`scale`).

Calculation of DREAM activity. DREAM-regulated genes (n = 328 genes) were defined as genes with promoters concurrently bound by at least 3 core DREAM complex subunits in a ChIP-seq experiment targeting LIN9, LIN54, p130, and E2F4 in quiescent human cells [24]. The Single Sample Gene-Set Enrichment (ssGSEA [78]) of the expression of these genes was calculated in each RNA-seq or scRNA-seq sample. For datasets in which the expression of one or more DREAM regulated genes was not detected, the remaining genes were used in the ssGSEA calculation. The ssGSEA "normalized enrichment score" was regressed against the number of expressed genes and total count of reads in each cell/sample, to correct for technical variation arising from sequencing depth [79]. The residuals from this regression (i.e., the normalized enrichment score corrected for sequencing depth) were inverted and scaled to be positive, making higher scores correspond to greater DREAM activity (i.e., lower expression of DREAM-regulated genes). The resulting value was termed "DREAM activity" (**Fig. 1**).

Proteomic DREAM activity validation. Matched transcriptomic and proteomic data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) [80] were used to compare the transcriptomic DREAM activity score with the abundance and phosphorylation state of LIN52 and p107 (RBL1). A total of 160 individuals were included from the clear cell renal cell carcinoma

(ccRCC) and head and neck squamous cell carcinoma (HNSCC) cohorts; samples in which the DREAM components were not detected were excluded. Multivariate linear regression was performed using the model:

$$DREAM\ activity \sim tissue\ type\ +\ LIN52\ +\ p107\ +\ LIN52_{s28\ phosphorylation} + p107_{phosphorylation}$$

where $p107_{phosphorylation}$ was the mean phosphorylation across 17 detected phospho-sites on p107.

Identification of somatic mutations in scRNA-seq. For the Tabula Sapiens [42] and SEA-AD [81] sc/snRNA-seq datasets, SComatic [41] was applied with default parameters to identify somatic single nucleotide substitution mutations. In brief, first somatic mutations in each "pseudo-bulked" cell type in each individual were identified, treating variants present in more than one cell type as germline. Second, somatic mutations in each individual cell were identified and filtered to include only variants also identified at the "pseudo-bulk" level of the respective cell type. Third, the count of somatic mutations in each cell was divided by the number of basepairs from which mutations were called in that cell – due to there being sufficient coverage at these loci – to define the somatic mutation rate per base pair. Cells without sufficient coverage for mutation identification in greater than 10 kb were discarded. The approximate number of somatic mutations per cell was then calculated by multiplying the somatic mutation rate per base pair by the approximate number of base pairs in the haploid human genome (3.0×10^9). Finally, the somatic mutation rate per cell per year was calculated by dividing the number of somatic mutations per cell by the age of the donor of that cell. For the Tabula Muris Senis dataset, somatic mutations had

been identified by the original publication [33], and we simply scaled these somatic mutation burdens to be relative to the number of bases from which mutations were called.

Somatic mutation rate validation. The somatic mutation rate per kb per month, estimated from scRNA-seq in mouse skin cells (**Fig. 2**), was compared with a previous estimate of somatic mutation rates in mouse fibroblasts [7] obtained using the single-cell multiple displacement amplification (SCMDA) DNA-sequencing method [82]. Milholland et al. [7] found there to be approximately 4.4×10^{-7} somatic mutations per bp in a 5 day old mouse. We converted this estimate to be in units of mutations/kb/month, using the average month length of 30.44 days, and found our estimate of 0.0030 mutations/kb/month to be remarkably similar to this previous estimate of 0.0027 mutations/kb/month.

$$\frac{4.4\times10^{-7}\ mutations}{5\ days\times 1\ bp}\times\frac{30.437\ days}{1\ month}\times\frac{1{,}000\ bp}{1\ kb}=\ 0.0027$$

We also compared the mutation rates we identified from human sc/snRNA-sequencing (**Fig. 5**) to previous estimates based on DNA sequencing of the same or similar cell types as those considered in our analyses. This included seven estimates of mutation rates in immune cells [83,84], three in neurons [83,85], and one in each of muscle satellite cells, skin fibroblasts, and cells from the bladder urothelium [83]. DNA- and RNA-based mutation estimates were generally consistent though the rates we now observed were marginally lower (**Supplementary Fig. 6b-f**). This difference in mutation rates may reflect that RNA, unlike previous measures from DNA, is primarily derived from coding sequences which have reduced somatic mutation rates relative to non-coding sequences due to increased mismatch and transcription-coupled repair [8,86–88].

Generation of random background distributions. To serve as background distributions for the association of DREAM activity with somatic mutation rate (**Fig. 2g-h**), species maximum lifespan (**Fig. 4c**), and age at AD onset (**Supplementary Fig. 6h-i**), the association between the expression of random gene sets and these measures was calculated. In each dataset, 500 – 1,000 sets of genes were randomly selected, with replacement, from all genes detected in that dataset. The number of genes in each of these random sets was defined to be the same as the number of DREAM regulated genes detected in that dataset (n = 249 for the TMS mouse scRNA-seq, n = 308 for the cross-species RNA-seq dataset, and n = 322 for the ROSMAP human brain RNA-seq). The "activity" of these random gene sets was then calculated by treating these genes as if they were DREAM regulated genes and applying the process that was used to calculate DREAM activity (i.e., ssGSEA followed by sequencing depth correction and inversion, see the 'Calculation of DREAM activity' methods section). Finally, the activity of these random gene sets was compared to the respective measures of interest and the strength of that association was contrasted to that of DREAM activity and the measure of interest.

Comparison of lifespan across species. Cross-species comparisons of DREAM activity, maximum lifespan, adult weight, and somatic mutation rate were conducted using a dataset of gene expression data (RNA-seq) collected from 3 tissues (brain, liver, kidney) of 92 different species [35] (n = ~3 samples per species per tissue, n = 803 total samples). We first calculated a representative DREAM activity for each species in each tissue by averaging the DREAM activity scores of all individuals of a given species in a tissue. Species with fewer than two sampled individuals in a given tissue were discarded from comparisons considering that tissue; taxonomic orders with only a single sampled species were discarded. We examined the association of DREAM activity with maximum lifespan (ML) using zero-intercept reduced major axis regression

(RMA, implemented by the pylr2 regress2 function [89]), in each tissue separately (**Fig. 4b, c**). Because body size can confound interspecies lifespan relationships [90], we conducted the same DREAM vs. ML comparison but correcting for adult body weight (AW). To do so, we performed an allometric regression of each trait (DREAM activity and ML) on $\log_{10}$-transformed AW and compared the resulting residuals (i.e., partial regression of DREAM activity vs. ML with respect to AW, **Fig. 5d, Supplementary Fig. 3a-b**). Finally, integrating a previously published estimate of somatic mutation rates [4] for a subset of the same species, we performed analogous cross-species comparisons of DREAM complex activity and somatic mutation rate, both with and without adjusting for AW (**Fig. 4f**, **Supplementary Fig. 3c-g**).

Survival analysis of mouse strains. The Williams et al. [40] dataset consists of 50 distinct strains of inbred mice split into two cohorts: the "Deeply Profiled Cohort," which was profiled for liver gene expression (RNA-seq), blood biomarkers, and the weight of different organs (n = 270 mice); and the "Lifespan Cohort," which was followed to measure the lifespan of each strain (n = 612 mice). The DREAM activity of each mouse in the Deeply Profiled Cohort was calculated from the liver RNA-seq. Four strains with highly variable DREAM activity scores across mice were removed (standard deviation > 0.2). The Deeply Profiled Cohort was used to estimate strain-specific covariates for use in a Cox proportional hazards model (lifelines CoxPHFitter [91]). This model was fit to the survival data of the Lifespan Cohort to measure the association of lifespan with diet (high fat vs. chow diet), DREAM activity, weight, clinical biomarkers, and weight change (**Fig. 4**).

$$survival \sim diet \times (DREAM\ activity + PC1_{weight\ \&\ blood\ lipids} + seq.depth + \Delta weight)$$

Covariates in the Cox model included: the mean DREAM activity value within each strain on each diet; the mean sequencing depth in each strain, to account for any confounding of DREAM activity

by differences in sequencing depth between strains; the first principal component (PC1, 25.2% of variance explained, **Supplementary Fig. 5g**) from an analysis of normalized clinical biomarkers and organ weights (variables listed in **Supplementary Fig. 5g**); diet group (chow or high fat diet); and change in body weight since starting the respective diet.

Neuropathology vs. DREAM activity. Neurodegenerative pathology was compared with DREAM complex activity using the SEA-AD [81] dataset, in which snRNA-seq was carried out on 803,112 brain cells from 83 individuals. Neuropathological assessments, including Thal phase, CERAD score, and Braak stage, had been conducted for each individual and integrated into an overall AD neuropathological change score [92] (ADNC). Each variable was binarized into high (Thal ≥ 3, Cerad ≤ 2, Braak ≥ 4, ADNC ≥ 2) versus low AD risk. Then, linear mixed effect regression was used to test the association between each of these variables – in addition to APOE 4/4 genotype, age (scaled to units of decades), and the count of somatic mutations per cell (log2 scaled) – and DREAM activity (Fig. 5e). These variables were treated as fixed effects, while the individual identifier was modeled as a random effect, to account for the nested structure of the data, with multiple cells sampled per individual.

Association of DREAM activity with age at AD diagnosis. The ROSMAP [44] (Religious Orders Study and Rush Memory and Aging Project) dataset was used to investigate the association of DREAM activity with the diagnosis of Alzheimer's disease. In this study, RNA-sequencing data were collected from multiple brain regions (dorsolateral prefrontal cortex, head of the caudate nucleus, posterior cingulate cortex, temporal cortex, and frontal cortex) of deceased donors (mean 2.4 ± std 1.0 brain regions per individual, n = 1,105 individuals, n = 2,776 samples). A Cox proportional hazard model (lifelines CoxPHFitter [91]) was fit to predict the age at AD diagnosis of these individuals. Age at AD diagnosis was right-censored at 90 years to ensure masking of

donor identity. The DREAM activity, number of APOE 4 alleles, sex, total RNA sequencing depth, count of genes with detected expression, and the duration of time between death and tissue procurement (post-mortem interval) were used as covariates in the model. DREAM activity scores were Z-score normalized within each brain region and outliers were dropped (samples with DREAM $|Z| > 5$). As there were multiple samples from each individual, the model was corrected for intra-individual correlations through clustered variance adjustment (**Fig. 5f-g**).

DREAM deficient mice. We leveraged a previous study [45] which had engineered mice with inactivated DREAM activity. In these mice, DREAM assembly was prevented by combining a constitutive *Rbl1* missense allele ($p107^{D/D}$), which disrupts MuvB binding, with tamoxifen-induced UBC-Cre-ERT2 deletion of *Rbl2*. Because either p107 or p130 can scaffold DREAM, loss of p130 in the $p107^{D/D}$ background abolishes assembly of the complex [45]. Control mice were Cre-negative. At 8 weeks of age, both groups were treated with tamoxifen – this activated the deletion of gene encoding for p130 (*Rbl2*) producing DREAM loss-of-function (LoF) mice ($p107^{D/D}$ $p130^{-/-}$, n = 5), while the control mice retained intact DREAM activity due to Cre-negativity ($p107^{D/D}$ $p130^{fl/fl}$, n = 5). Brain tissue was obtained from these mice at death, formalin-fixed, and paraffin-embedded.

Duplex sequencing (UDSeq) library preparation and sequencing. Formalin-fixed paraffin-embedded (FFPE) mouse brain tissues (n = 10) were sectioned, stained with hematoxylin and eosin, and punches containing sufficient nuclei were selected by microscopic inspection. Genomic DNA was extracted using the QIAamp DNA FFPE Advanced Kit (Qiagen, cat. no. 56704) according to the manufacturer's instructions and quantified by Qubit, with 500 ng per sample normalized to 5 ng/µl for library preparation. Genomic DNA was enzymatically fragmented using the UltraShear system (M7634L) for 25 minutes, targeting an average fragment size of

approximately 350 base pairs. The fragmented DNA was then processed using the UDSeq library preparation workflow with the xGen™ cfDNA & FFPE DNA Library Preparation Kit. All steps were performed on magnetic beads to minimize DNA loss during purification and enhance library conversion efficiency. For final library construction, 0.175 femtomoles of UMI-ligated DNA were amplified via 15 cycles of PCR to incorporate Illumina®-compatible dual index sequences. For matched normal samples, 5 femtomoles of UMI-ligated DNA from the same source were amplified using 10 PCR cycles. The resulting universal dual-indexed libraries were sequenced at the UCSD IGM Genomics Center using a 25B flow cell on the NovaSeq X platform.

Identification of somatic mutations in duplex sequencing data. Sequencing data were downloaded and analyzed within the Triton Shared Computing Cluster at the San Diego Supercomputer Center (https://doi.org/10.57873/T34W2R). Somatic mutation calling and mutational burden analysis were performed using DupCaller 1 (v1.0.1) [93] on UDSeq data with matched normal samples. Variant Call Format (VCF) files generated by DupCaller were used for mutational profiling and signature assignment, following our previously established methodology using the SigProfiler suite of tools. In particular, SigProfilerAssignment was run on the variant call set, with the entire mouse COSMIC catalog and FFPE signature from Guo et al. 2022 [47] FFPE signature as the input signatures. Mutations classified as FFPE artifacts were discarded and remaining mutations were analyzed. For each sample, the raw mutation count in the sequenced bases was extrapolated to estimate the number of mutations expected if the entire diploid genome had been covered at sufficient depth, yielding a per-cell equivalent mutation burden (termed "mutations per cell").

Analysis of UD-seq mutation calls. Negative binomial regression models were used to test for a difference in frequency of single base substitutions and insertion/deletion mutations between DREAM LoF and control mice. Models of three complexities were compared:

$$(Model\ 1)\ mutation\ count \sim DREAM\ status + genomic\ coverage + read\ families$$

$$(Model\ 2)\ mutation\ count \sim DREAM\ status + genomic\ coverage\ + read\ families + age$$

$$(Model\ 3)\ mutation\ count \sim DREAM\ status + genomic\ coverage \times read\ families + age$$

Model 3 showed the best fit across all mutation types as measured by log-likelihood (**Supplementary Fig. 7a-f**) and was therefore used for all statistical comparisons of mutation frequency. For comparisons of the proportional contribution of mutational processes, a binomial model of the same form was used. One outlier sample – identified by having a Cook's distance [94] > 1 and total mutation count |Z-score| > 3 across both mutation classes – was removed (**Supplementary Fig. 7g-i**).

Software. All analyses were performed in Python 3.10. 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. Specific statistical approaches used are noted in the respective methods sections and figure captions.

Data availability

Raw UD-sequencing data can be accessed on the Sequence Read Archive (SRA) at accession number [To be uploaded before publication]. Other data can be accessed through the respective publications (see **Supplementary Table 1**).

Code availability

All custom algorithms and analysis code are available in the GitHub repository at https://github.com/zanekoch/dream_proj.

Author Contributions

ZK designed the study, carried out the primary data analyses, and wrote the manuscript. ABD, DHM, BS, and LA designed the study and wrote the manuscript. FD and PP designed the study and provided the mouse samples. KL and SN designed the study, performed the duplex sequencing experiment, and wrote the manuscript. TI designed the study and wrote the manuscript.

Declaration of Interests

TI is a co-founder, member of the advisory board, and has an equity interest in Data4Cure and Serinus Biosciences. TI is a consultant for and has an equity interest in Ideaya Biosciences and Eikon Pharmaceuticals. The terms of these arrangements have been reviewed and approved by the University of California San Diego in accordance with its conflict-of-interest policies. L.B.A. is a co-founder, CSO, scientific advisory member, and consultant for Acurion (formerly io9), has equity and receives income. The terms of this arrangement have been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. L.B.A. is also a compensated member of the scientific advisory board of Inocras. L.B.A.'s spouse is an employee of Hologic, Inc. L.B.A. declares U.S. provisional applications filed with UCSD with serial numbers: 63/269,033; 63/289,601; 63/483,237; 63/412,835; 63/492,348; and 63/366,392 as well as a European patent application with application number EP25305077.7. L.B.A. and S.P.N. also declare provisional patent application PCT/US2023/010679. L.B.A. is also

an inventor of a US Patent 10,776,718 for source identification by non-negative matrix factorization. All other authors declare no conflicts of interest.

Figures

Figure 9

a) The DREAM transcriptional repressor complex is formed in G0 following the phosphorylation of MuvB by DYRK1A. Upon reentry to the cell cycle, the DREAM complex dissociates and DREAM-target genes can be expressed. Here we explore the relationship between DREAM activity and somatic mutation, as well as between DREAM activity and rate of aging.

b) Calculation of DREAM activity from target gene expression. Single sample gene set enrichment analysis (ssGSEA) is used to summarize the mRNA expression values of all detected DREAM complex target genes (columns, 328 DREAM-regulated genes) in each cell (rows). This ssGSEA score is then normalized and corrected for sequencing depth to produce a per cell DREAM activity score (**Methods**).

c) U2OS cells are exposed to a DREAM complex inhibitor (harmine, INDY) versus control treatment (harmine control, INDY control). mRNA expression profiles are measured, from which we compute DREAM activity. Bar heights indicate mean DREAM activity across samples (n = 4 biological replicates), error bars denote 95% confidence intervals, and p-values indicate the result of two-sided t-tests.

d) Similar to (c), but showing DREAM complex activity in wild type (WT) vs. *lin-52*(*n771*) mutant *C. elegans*.

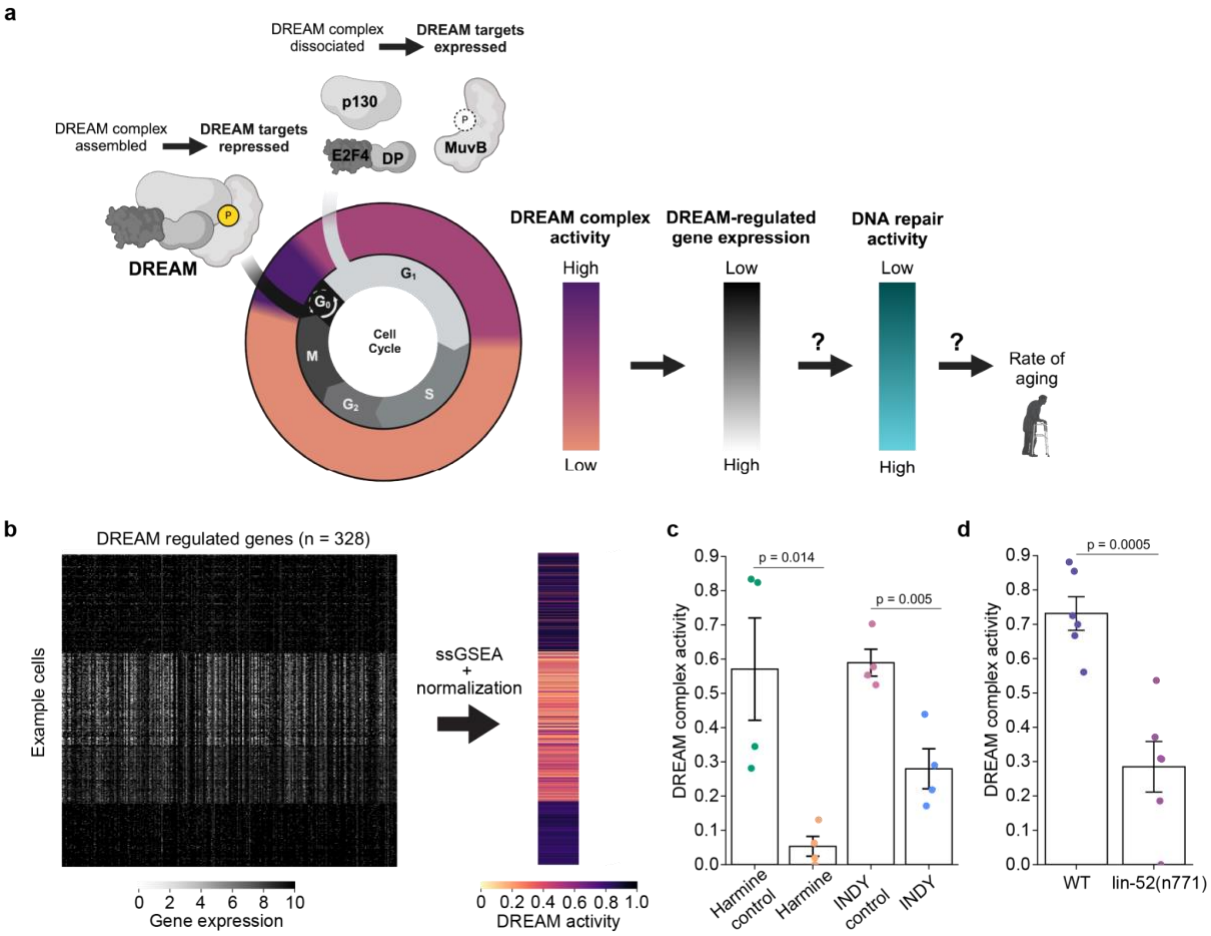

a
DREAM complex dissociated
DREAM targets expressed
DREAM complex assembled
DREAM targets repressed
p130
E2F4
DP
MuvB
DREAM
Cell Cycle
DREAM complex activity
DREAM-regulated gene expression
DNA repair activity
Rate of aging
b
DREAM regulated genes (n = 328)
Example cells
ssGSEA + normalization
Gene expression
DREAM activity
c
DREAM complex activity
p = 0.014
p = 0.005
Harmine control
Harmine
INDY control
INDY
d
p = 0.0005
WT
lin-52(n771)

Figure 10

a) Box plots of the distribution of somatic mutation burden across the cells of each tissue at each age (n = 110,824 cells, n = 18 mice). Mutation burden is expressed as the mutation count per callable kilobase (**Methods**). (***) and (n.s.) indicate p values calculated by modeling Spearman ρ's as a Student's t distribution, yielding $p < 5.0 \times 10\text{-}19$ and $p \geq 0.01$, respectively. BAT, GAT, MAT, and SCAT abbreviate brown, gonadal, mesenteric, and subcutaneous adipose tissue, respectively. Boxes show inter-quartile range (IQR) with median line; whiskers extend to 1.5 IQR.
b) The somatic mutation burden per kilobase in cells from 18 month old mice (n = 34,027 cells, n = 4 mice) stratified by DREAM complex activity (highest vs. lowest 20% of DREAM complex activities within each tissue). Error bars denote 95% confidence intervals. (*) indicates $p < 0.0023$ (Bonferroni corrected p-value) and (n.s.) indicates $p \geq 0.0023$, based on a two-sided Mann-Whitney test.
c) Uniform manifold approximation and projection (UMAP) of all cells of all age mice from the first 12 tissue types in (b) (from the left, n = 42,508 cells, n = 18 mice), colored by cell type and labeled with tissue of origin.
d) As in (c) but coloring cells by DREAM activity. Cells with the 50% lowest and 50% highest DREAM activities for age and tissue type are colored orange and purple, respectively.
e) As in (d) but coloring cells by somatic mutation burden per kb. Cells with the 50% lowest and 50% highest somatic mutation burdens for their age and tissue type are colored green and pink, respectively.
f) Violin plot of the somatic mutation burden in hepatocyte cells (n = 1,162 cells, n = 18 mice) stratified by age and DREAM complex activity. (***) indicates a significant ($p < 5.0 \times 10^{-200}$) Spearman correlation between DREAM complex activity and mutation burden within an age group. Spearman ρ: 3 months = 0.16, 18 months = 0.30, and 24 months = 0.21.
g) Kernel density plot depicting the distribution of Spearman ρ's computed in each cell type (n = 116 cell types, n = 34,027 cells from 18 month old mice, n = 4 mice) between somatic mutation burden and gene-set activity scores (ssGSEA, **Methods**). Purple: DREAM-regulated gene set (n = 248 genes). Gray: 500 random gene sets (n = 248 random genes per set). (***) indicates a significant ($p < 4.86 \times 10^{-14}$) difference in Spearman correlation values between groups based on a two-sided Mann Whitney U test.
h) Similar to (g), but comparing somatic mutation burden and gene expression within *tissue* types (n = 21 tissue types, n = 34,027 cells from 18 month old mice, n = 4 mice, **Methods**).

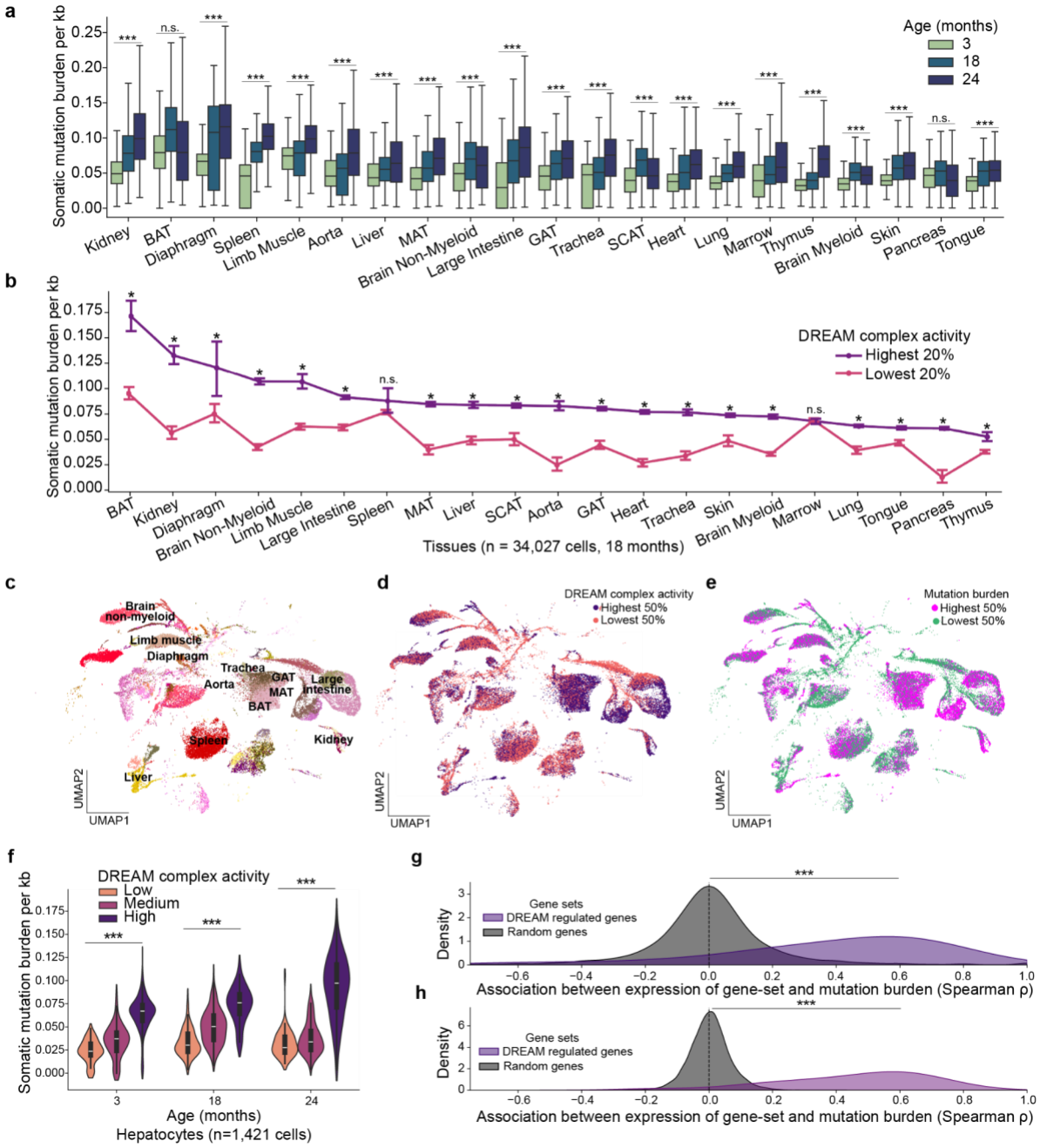

a
Somatic mutation burden per kb
Age (months)
3
18
24
Kidney
BAT
Diaphragm
Spleen
Limb Muscle
Aorta
Liver
MAT
Brain Non-Myeloid
Large Intestine
GAT
Trachea
SCAT
Heart
Lung
Marrow
Thymus
Brain Myeloid
Skin
Pancreas
Tongue
b
Somatic mutation burden per kb
DREAM complex activity
Highest 20%
Lowest 20%
Tissues (n = 34,027 cells, 18 months)
c
Brain non-myeloid
Limb muscle
Diaphragm
Trachea
Aorta
GAT
MAT
Large Intestine
BAT
Spleen
Kidney
Liver
UMAP1
UMAP2
d
DREAM complex activity
Highest 50%
Lowest 50%
e
Mutation burden
Highest 50%
Lowest 50%
f
DREAM complex activity
Low
Medium
High
Somatic mutation burden per kb
Age (months)
Hepatocytes (n=1,421 cells)
g
Gene sets
DREAM regulated genes
Random genes
Density
Association between expression of gene-set and mutation burden (Spearman ρ)
h
Gene sets
DREAM regulated genes
Random genes
Density
Association between expression of gene-set and mutation burden (Spearman ρ)

Figure 11

a) Barplots indicating the average DREAM activity across tissues (liver, kidney, brain), 1/maximum lifespan, and $\log_{10}$ adult weight in grams for each species (n = 92 species and 803 samples). Within each taxonomic order, species are sorted by decreasing average DREAM complex activity.
b) Scatterplot of maximum lifespan vs. DREAM activity measured in the liver (n = 92 species and 241 samples, **Methods**). Points indicate the mean value of each species, while the orange curve shows reduced major axis regression line fit to the actual values of all samples. The inner shaded area indicates one standard deviation from the regression line and the outer the 95% confidence interval.
c) Barplots indicating the mean and 99% confidence interval of the Spearman correlation between 1/lifespan and average DREAM activity of each species, in each tissue, iteratively leaving one species out (n = 92 species). (***) indicates $p < 8.23 \times 10^{-24}$ based on a two-sided Mann-Whitney test.
d) Scatterplot and reduced major axis regression depicting the average adult weight (AW) adjusted DREAM activity and lifespan of the species in (a) (n = 92 species and 241 liver samples, **Methods**). Both axes are shown on a logarithmic scale.
e) Heatmap of the Spearman correlation coefficients comparing DREAM activity scores between each pair of tissues, across the species and samples in (a). All pairwise correlations are significant ($p < 0.0005$).
f) Scatterplot and reduced major axis regression depicting the average AW adjusted DREAM activity and somatic mutation rate (in the liver and colon, respectively) across species (n = 11 species, **Methods**).

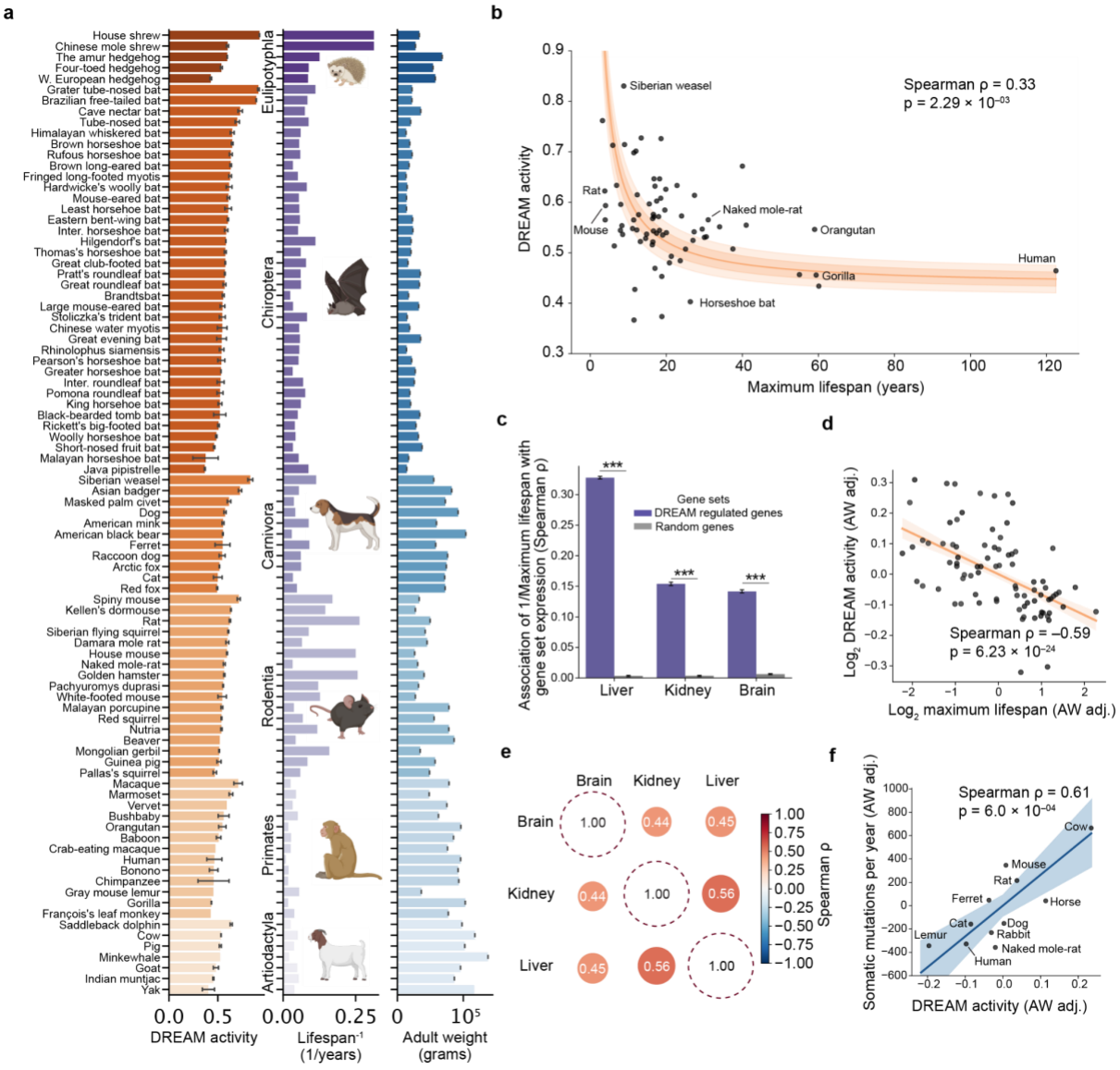

a
House shrew
Chinese mole shrew
The amur hedgehog
Four-toed hedgehog
W. European hedgehog
Grater tube-nosed bat
Brazilian free-tailed bat
Cave nectar bat
Tube-nosed bat
Himalayan whiskered bat
Brown horseshoe bat
Rufous horseshoe bat
Brown long-eared bat
Fringed long-footed myotis
Hardwicke's woolly bat
Mouse-eared bat
Least horseshoe bat
Eastern bent-wing bat
Inter. horseshoe bat
Hilgendorf's bat
Thomas's horseshoe bat
Great club-footed bat
Pratt's roundleaf bat
Great roundleaf bat
Brandtsbat
Large mouse-eared bat
Stoliczka's trident bat
Chinese water myotis
Great evening bat
Rhinolophus siamensis
Pearson's horseshoe bat
Greater horseshoe bat
Inter. roundleaf bat
Pomona roundleaf bat
King horsehoe bat
Black-bearded tomb bat
Rickett's big-footed bat
Woolly horseshoe bat
Short-nosed fruit bat
Malayan horseshoe bat
Java pipistrelle
Siberian weasel
Asian badger
Masked palm civet
Dog
American mink
American black bear
Ferret
Raccoon dog
Arctic fox
Cat
Red fox
Spiny mouse
Kellen's dormouse
Rat
Siberian flying squirrel
Damara mole rat
House mouse
Naked mole-rat
Golden hamster
Pachyuromys duprasi
White-footed mouse
Malayan porcupine
Red squirrel
Nutria
Beaver
Mongolian gerbil
Guinea pig
Pallas's squirrel
Macaque
Marmoset
Vervet
Bushbaby
Orangutan
Baboon
Crab-eating macaque
Human
Bonono
Chimpanzee
Gray mouse lemur
Gorilla
François's leaf monkey
Saddleback dolphin
Cow
Pig
Minkewhale
Goat
Indian muntjac
Yak
Eulipotyphla
Chiroptera
Carnivora
Rodentia
Primates
Artiodactyla
0.0 0.5
DREAM activity
0.00 0.25
Lifespan-1 (1/years)
0 10^5
Adult weight (grams)
b
Siberian weasel
Spearman ρ = 0.33
p = 2.29 × 10−03
Rat
Mouse
Naked mole-rat
Orangutan
Human
Gorilla
Horseshoe bat
DREAM activity
Maximum lifespan (years)
c
Association of 1/Maximum lifespan with gene set expression (Spearman ρ)
Gene sets
DREAM regulated genes
Random genes
***
Liver
Kidney
Brain
d
Log2 DREAM activity (AW adj.)
Spearman ρ = −0.59
p = 6.23 × 10−24
Log2 maximum lifespan (AW adj.)
e
Brain
Kidney
Liver
1.00
0.44
0.45
0.56
Spearman ρ
f
Somatic mutations per year (AW adj.)
Spearman ρ = 0.61
p = 6.0 × 10−04
Cow
Mouse
Rat
Ferret
Horse
Cat
Dog
Rabbit
Lemur
Naked mole-rat
Human
DREAM activity (AW adj.)

Figure 12

a) Kaplan-Meier survival curve of high fat diet (HFD)-fed mice belonging to strains with the highest vs. lowest 20% of DREAM complex activities (purple vs. pink, n = 119 mice, n = 14 strains). The dark line and shaded area indicate the proportion of mice surviving to a particular age and the 95% confidence interval of this survival estimate, respectively. Dashed lines indicate the median survival of each group. P value calculated using a two-sided log-rank test.
b) Dot-and-whisker plot indicating the association of each variable on the y-axis with mouse survival, for mice fed the HFD (green, n = 299 mice) or chow diet (CD) (grey, n = 313 mice). Dots and whiskers indicate the hazard ratio and 95% confidence intervals from a Cox proportional hazard model (**Methods**). (***), (**), and (*) indicate p values, from a two-sided Wald test, of less than 5.0×10^{-6}, 0.005, and 0.05, respectively.
c) Scatterplot of mean DREAM activity of each mouse strain on the CD (x axis), compared to the change in survival when members of that strain are placed on the HFD. n = 29 strains profiled on both diets; n = 10 mice per diet per strain on average. P value calculated by modeling Spearman ρ's as a Student's t distribution.

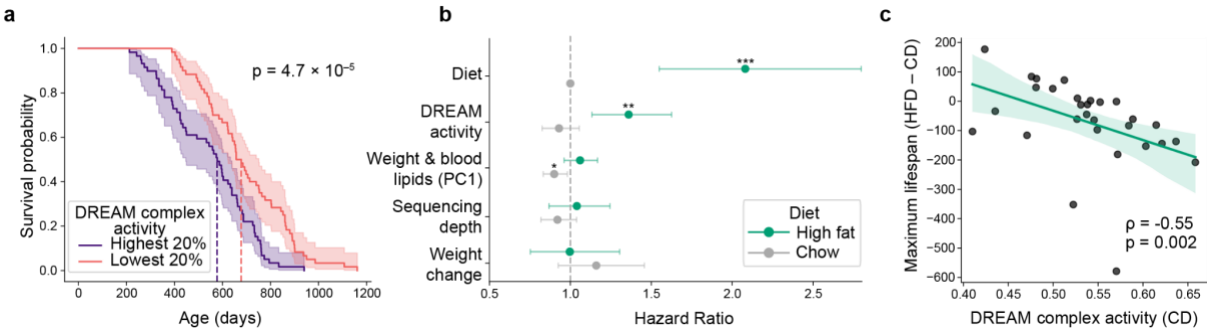

a
b
c
p = 4.7 × 10⁻⁵
Survival probability
Age (days)
DREAM complex activity
Highest 20%
Lowest 20%
Diet
DREAM activity
Weight & blood lipids (PC1)
Sequencing depth
Weight change
Hazard Ratio
Diet
High fat
Chow
Maximum lifespan (HFD – CD)
DREAM complex activity (CD)
ρ = -0.55
p = 0.002

Figure 13

a) 90 human individuals (80 with AD or neurodegeneration and 10 cognitively normal) were profiled for mRNA expression in single cells from 11 tissues (n = 803,122 total cells, 73% of cells were from the brain) as well as neuropathology and cognitive function. Cell type and somatic mutation profile of each cell were identified from the single-cell mRNA.
b) Scatter plot of the number of somatic mutations per haploid genome (Methods) identified in the brain of each individual, compared to the age of that individual (n = 77 individuals with ≥ 5,000 sequenced brain cells, n = 589,206 cells). Points and error bars indicate the mean and 95% confidence interval across all cells from an individual. P value calculated by modeling Spearman ρ's as a Student's t distribution.
c) Barplot indicating the mean and 95% confidence interval of the number of somatic mutations identified in each cell per year within each tissue type (normalized to haploid genome size, n = 803,122 cells, n = 90 individuals).
d) Number of somatic mutations in each cell (normalized to haploid genome size, n = 803,122 cells, n = 90 individuals, Methods) stratified by DREAM complex activity (highest vs. lowest 20% of DREAM activity scores within each tissue). Error bars denote 95% confidence intervals. (***), (*), and (n.s.) indicate p values from a one-sided Mann-Whitney test of $p < 5.0 \times 10\text{-}17$, $p < 0.005$ (Bonferroni corrected p-value), and $p \geq 0.005$, respectively.
e) Association of cell or individual-level covariates with DREAM complex activity in neurons (n = 486,912 neurons, n = 77 individuals). Dots and whiskers indicate the coefficient and 95% confidence interval from linear mixed effects modelling (**Methods**). (***), (**), and (*) indicate p values from a two-sided t-test of $p < 5.0 \times 10^{-30}$, $p < 5.0 \times 10^{-4}$, and $p < 0.05$, respectively. Covariates with insignificant associations are shown in grey.
f) Kaplan-Meier survival curve comparing the age at AD diagnosis between individuals with the top and bottom third of DREAM complex activities (purple vs. pink, n = 1,101 individuals). The dark line and shaded area indicate the proportion of individuals without AD at a particular age and the 95% confidence interval of this estimate, respectively. P value calculated using a two-sided log-rank test.
g) Dot-and-whisker plot indicating the association of each variable on the y-axis with age at AD diagnosis (n = 1,101 individuals). Dots and whiskers indicate the hazard ratio and 95% confidence intervals from a Cox proportional hazard model (**Methods**). Asterisks indicate p values from a two-sided Wald test; (***), $p < 5.0 \times 10^{-10}$ and (**), $p < 0.01$. Covariates with insignificant associations are shown in grey. Inset: Line plot of the predicted effect of different levels of DREAM on age at AD diagnosis, holding other factors constant.

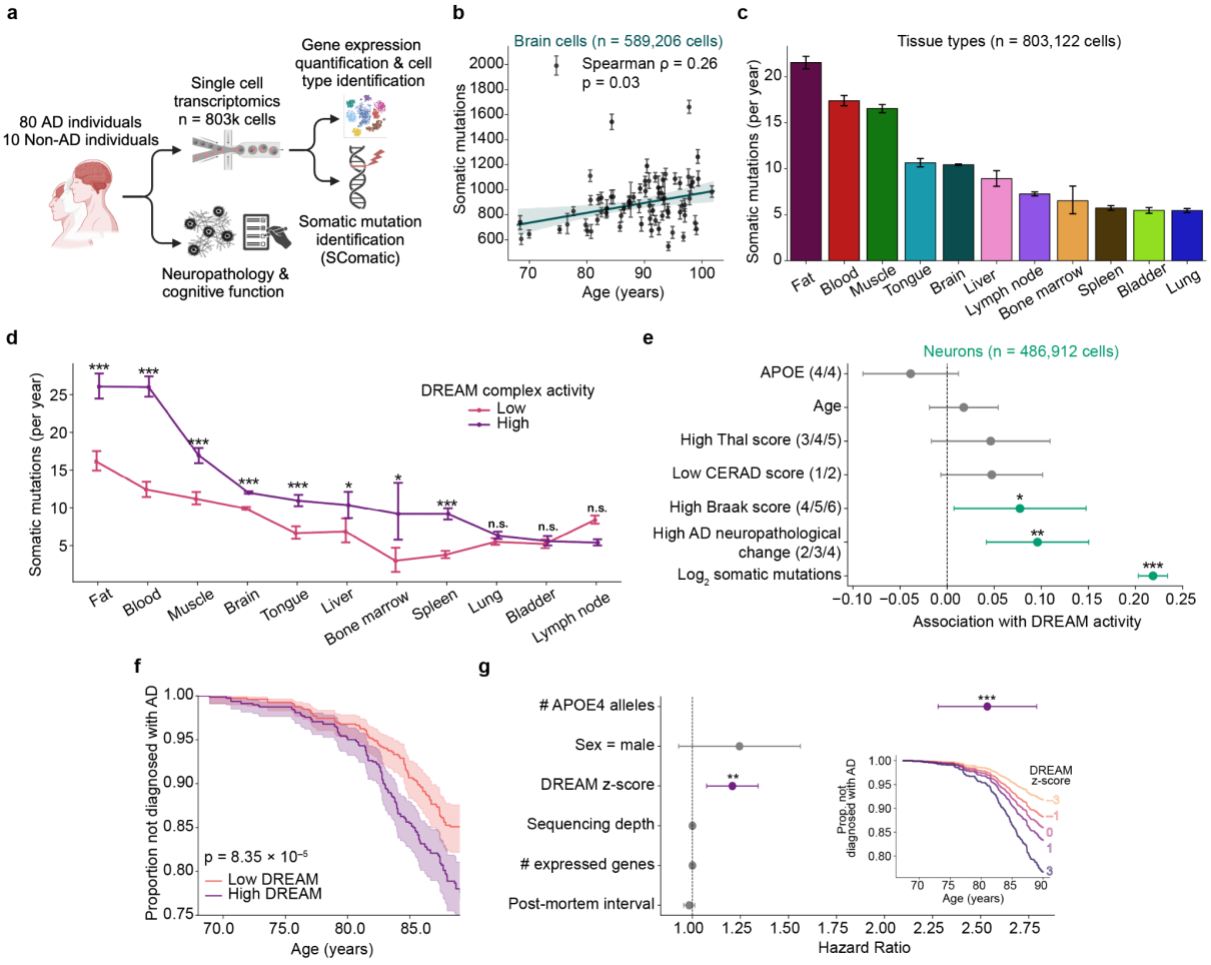

a
80 AD individuals
10 Non-AD individuals
Single cell transcriptomics n = 803k cells
Gene expression quantification & cell type identification
Somatic mutation identification (SComatic)
Neuropathology & cognitive function
b
Brain cells (n = 589,206 cells)
Spearman ρ = 0.26
p = 0.03
Somatic mutations
Age (years)
c
Tissue types (n = 803,122 cells)
Somatic mutations (per year)
Fat
Blood
Muscle
Tongue
Brain
Liver
Lymph node
Bone marrow
Spleen
Bladder
Lung
d
Somatic mutations (per year)
DREAM complex activity
Low
High
n.s.
e
Neurons (n = 486,912 cells)
APOE (4/4)
Age
High Thal score (3/4/5)
Low CERAD score (1/2)
High Braak score (4/5/6)
High AD neuropathological change (2/3/4)
Log2 somatic mutations
Association with DREAM activity
f
Proportion not diagnosed with AD
p = 8.35 × 10−5
Low DREAM
High DREAM
Age (years)
g
# APOE4 alleles
Sex = male
DREAM z-score
Sequencing depth
expressed genes
Post-mortem interval
Hazard Ratio
Prop. not diagnosed with AD
Age (years)

Figure 14

a) Whole brain tissue slices were collected from control mice ($p107^{D/D}$ $p130^{fl/fl}$, n = 5) or DREAM loss-of-function mice (LoF, $p107^{D/D}$ $p130^{-/-}$, n = 4). Somatic mutations were identified via duplex sequencing (UDseq) and somatic variant calling (DupCaller).

b) Single base substitution (SBS) mutations per (diploid) cell in each mouse. Points are individuals; diamonds and whiskers denote the means and 95% confidence intervals (CI). P value reflects the association between DREAM status and the count of SBS mutations per cell according to a negative binomial regression model (**Methods**).

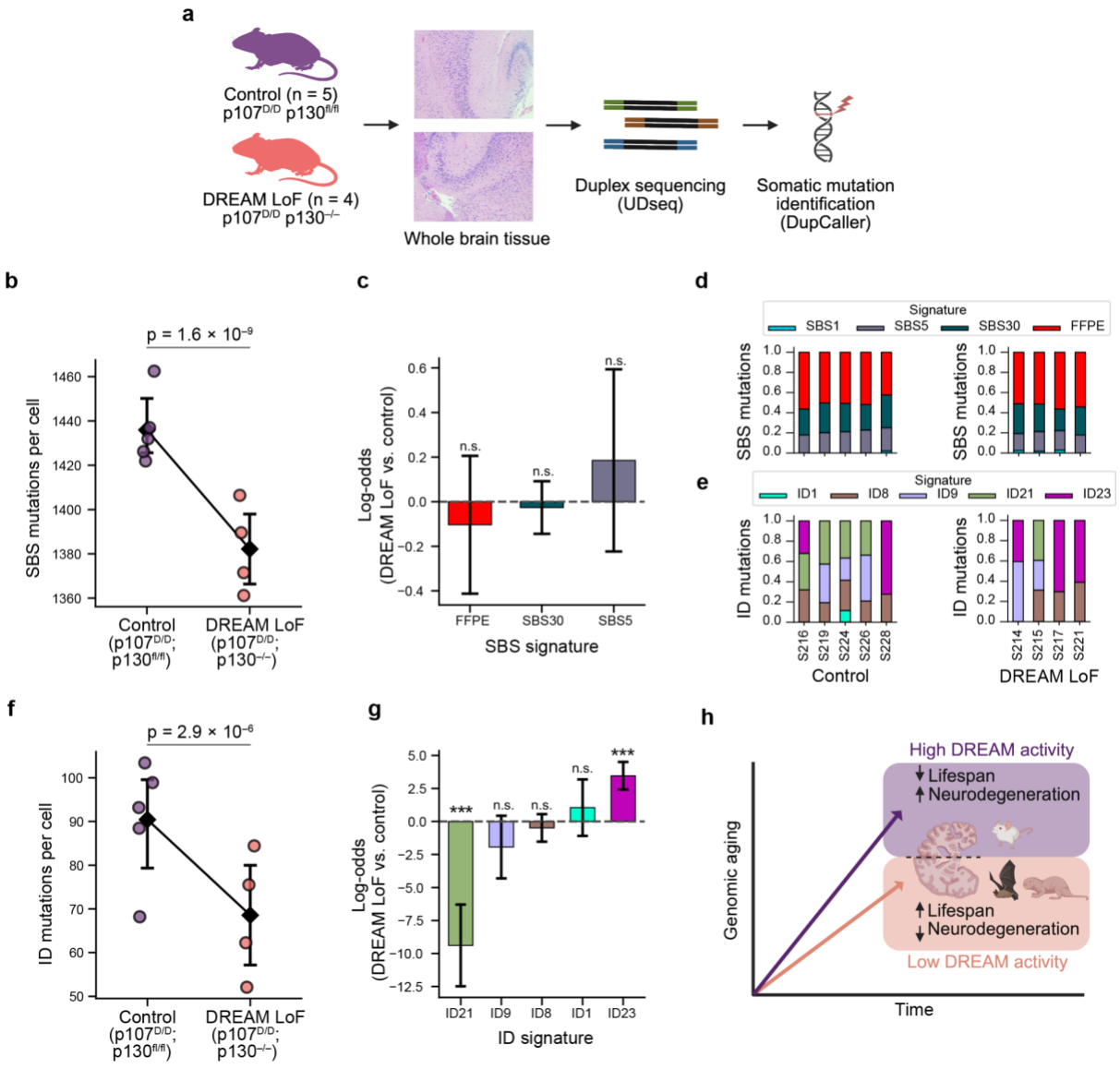
c) Log-odds of each SBS signature in DREAM LoF vs. control mice. Error bars indicate 95% CI. P values indicate significance of association between DREAM status and the proportion of mutations assigned to each signature from a two-sided binomial model (**Methods**), with n.s. corresponding to $p > 0.05$.

d-e) Proportional contributions of (d) SBS signatures and (e) ID signatures (colored stacked bars) in each control or DREAM LoF mouse (x axis). The tissue preservation process (FFPE, formalin-fixed paraffin-embedded) is associated with a distinct SBS mutational signature.

f) Similar to (b), but showing insertion/deletion (ID) mutations.

g) Similar to (c), but for insertion/deletion (ID) mutational signatures. (***) and (n.s.) indicate $p < 5.0 \times 10^{-9}$ and $p > 0.05$, respectively.

h) Model for the relationship between DREAM activity and aging phenotypes, in which higher DREAM activity associates with a faster rate of genomic aging due to repression of DNA repair. Consistent with this model, short lived species have higher DREAM activity than long lived species, and individuals with higher DREAM complex activity have accelerated neurodegeneration.


a
Control (n = 5)
p107D/D p130fl/fl
DREAM LoF (n = 4)
p107D/D p130−/−
Whole brain tissue
Duplex sequencing
(UDseq)
Somatic mutation
identification
(DupCaller)
b
p = 1.6 × 10−9
SBS mutations per cell
Control
(p107D/D;
p130fl/fl)
DREAM LoF
(p107D/D;
p130−/−)
c
Log-odds
(DREAM LoF vs. control)
n.s.
FFPE
SBS30
SBS5
SBS signature
d
Signature
SBS1
SBS5
SBS30
FFPE
SBS mutations
e
Signature
ID1
ID8
ID9
ID21
ID23
ID mutations
S216
S219
S224
S226
S228
S214
S215
S217
S221
Control
DREAM LoF
f
p = 2.9 × 10−6
ID mutations per cell
Control
(p107D/D;
p130fl/fl)
DREAM LoF
(p107D/D;
p130−/−)
g
Log-odds
(DREAM LoF vs. control)
***
n.s.
ID21
ID9
ID8
ID1
ID23
ID signature
h
High DREAM activity
Lifespan
Neurodegeneration
Lifespan
Neurodegeneration
Low DREAM activity
Genomic aging
Time

References

1. Kratz A, Kim M, Kelly MR, Zheng F, Koczor CA, Li J. A multi-scale map of protein assemblies in the DNA damage response. Cell Syst. 2023;14: 447–463.e8.

2. Harper JM, Salmon AB, Leiser SF, Galecki AT, Miller RA. Skin-derived fibroblasts from long-lived species are resistant to some, but not all, lethal stresses and to the mitochondrial inhibitor rotenone. Aging Cell. 2007;6: 1–13.

3. Ma S, Upneja A, Galecki A, Tsai Y-M, Burant CF, Raskind S. Cell culture-based profiling across mammals reveals DNA repair and metabolism as determinants of species longevity. Elife. 2016;5. doi:10.7554/eLife.19130

4. Cagan A, Baez-Ortega A, Brzozowska N, Abascal F, Coorens THH, Sanders MA. Somatic mutation rates scale with lifespan across mammals. Nature. 2022;604: 517–524.

5. Zhang L, Dong X, Tian X, Lee M, Ablaeva J, Firsanov D. Maintenance of genome sequence integrity in long- and short-lived rodent species. Sci Adv. 2021;7: eabj3284.

6. Tian X, Firsanov D, Zhang Z, Cheng Y, Luo L, Tombline G. SIRT6 Is Responsible for More Efficient DNA Double-Strand Break Repair in Long-Lived Species. Cell. 2019;177: 622–638.e22.

7. Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. Differences between germline and somatic mutation rates in humans and mice. Nat Commun. 2017;8: 15183.

8. Moore L, Cagan A, Coorens THH, Neville MDC, Sanghvi R, Sanders MA. The mutational landscape of human somatic and germline cells. Nature. 2021;597: 381–386.

9. Lynch M. Evolution of the mutation rate. Trends Genet. 2010;26: 345–352.

10. Sebastiani P, Solovieff N, Dewan AT, Walsh KM, Puca A, Hartley SW. Genetic signatures of

exceptional longevity in humans. PLoS One. 2012;7: e29848.

11. Gurinovich A, Song Z, Zhang W, Federico A, Monti S, Andersen SL. Effect of longevity genetic variants on the molecular aging rate. GeroScience. 2021;43: 1237–1251.

12. Simon M, Yang J, Gigas J, Earley EJ, Hillpot E, Zhang L. A rare human centenarian variant of SIRT6 enhances genome stability and interaction with Lamin A. EMBO J. 2022;41: e110393.

13. López-Otín C, Blasco MA, Partridge L, Serrano M, Kroemer G. Hallmarks of aging: An expanding universe. Cell. 2023;186: 243–278.

14. Schumacher B, Pothof J, Vijg J, Hoeijmakers JHJ. The central role of DNA damage in the ageing process. Nature. 2021;592: 695–703.

15. Arvanitaki ES, Goulielmaki E, Gkirtzimanaki K, Niotis G, Tsakani E, Nenedaki E. Microglia-derived extracellular vesicles trigger age-related neurodegeneration upon DNA damage. Proceedings of the National Academy of Sciences. 2024;121: e2317402121.

16. Yousefzadeh MJ, Huerta Guevara AP, Postmus AC, Flores RR, Sano T, Jurdzinski A. Failure to repair endogenous DNA damage in β-cells causes adult-onset diabetes in mice. Aging Biol. 2023;1. doi:10.59368/agingbio.20230015

17. Yoshimoto N, Hayashi K, Hishikawa A, Hashiguchi A, Nakamichi R, Sugita-Nishimura E. Significance of podocyte DNA damage and glomerular DNA methylation in CKD patients with proteinuria. Hypertens Res. 2023;46: 1000–1008.

18. Hishikawa A, Hayashi K, Yoshimoto N, Nakamichi R, Homma K, Itoh H. DNA damage and expression of DNA methylation modulators in urine-derived cells of patients with hypertension and diabetes. Sci Rep. 2020;10: 3377.

19. Nelson TJ, Xu Y. Sting and p53 DNA repair pathways are compromised in Alzheimer's disease. Sci

Rep. 2023;13: 8304.

20. Zhao H, Wang S, Li X. DNA damage accumulation in aging brain and its links to Alzheimer's disease progression. Genome Instability & Disease. 2022;3: 172–178.

21. Wyss-Coray T. Ageing, neurodegeneration and brain rejuvenation. Nature. 2016;539: 180–186.

22. Miller MB, Huang AY, Kim J, Zhou Z, Kirkham SL, Maury EA. Somatic genomic changes in single Alzheimer's disease neurons. Nature. 2022;604: 714–722.

23. Bujarrabal-Dueso A, Sendtner G, Meyer DH, Chatzinikolaou G, Stratigi K, Garinis GA. The DREAM complex functions as conserved master regulator of somatic DNA-repair capacities. Nat Struct Mol Biol. 2023;30: 475–488.

24. Litovchick L, Sadasivam S, Florens L, Zhu X, Swanson SK, Velmurugan S. Evolutionarily conserved multisubunit RBL2/p130 and E2F4 protein complex represses human cell cycle-dependent genes in quiescence. Mol Cell. 2007;26: 539–551.

25. Litovchick L, Florens LA, Swanson SK, Washburn MP, DeCaprio JA. DYRK1A protein kinase promotes quiescence and senescence through DREAM complex assembly. Genes Dev. 2011;25: 801–813.

26. Sadasivam S, DeCaprio JA. The DREAM complex: master coordinator of cell cycle-dependent gene expression. Nat Rev Cancer. 2013;13: 585–595.

27. Engeland K. Cell cycle arrest through indirect transcriptional repression by p53: I have a DREAM. Cell Death Differ. 2018;25: 114–132.

28. Guiley KZ, Liban TJ, Felthousen JG, Ramanan P, Litovchick L, Rubin SM. Structural mechanisms of DREAM complex assembly and regulation. Genes Dev. 2015;29: 961–974.

29. Kandhaya-Pillai R, Miro-Mur F, Alijotas-Reig J, Tchkonia T, Schwartz S, Kirkland JL. Key elements of cellular senescence involve transcriptional repression of mitotic and DNA repair genes through the p53-p16/RB-E2F-DREAM complex. Aging (Albany NY). 2023;15: 4012–4034.

30. Sit YT, Takasaki K, An HH, Xiao Y, Hurtz C, Gearhart PA. Synergistic roles of DYRK1A and GATA1 in trisomy 21 megakaryopoiesis. JCI Insight. 2023;8: e172851.

31. Murray A, Gough G, Cindrić A, Vučković F, Koschut D, Borelli V. Dose imbalance of DYRK1A kinase causes systemic progeroid status in Down syndrome by increasing the un-repaired DNA damage and reducing LaminB1 levels. EBioMedicine. 2023;94: 104692.

32. Uxa S, Castillo-Binder P, Kohler R, Stangner K, Müller GA, Engeland K. Ki-67 gene expression. Cell Death Differ. 2021;28: 3357–3370.

33. Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. Nature. 2020;583: 590–595.

34. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10: 1096–1098.

35. Liu W, Zhu P, Li M, Li Z, Yu Y, Liu G. Large-scale across species transcriptomic analysis identifies genetic selection signatures associated with longevity in mammals. EMBO J. 2023;42: e112740.

36. Yuzefovych LV, Musiyenko SI, Wilson GL, Rachek LI. Mitochondrial DNA damage and dysfunction, and oxidative stress are associated with endoplasmic reticulum stress, protein degradation and apoptosis in high fat diet-induced insulin resistance mice. PLoS One. 2013;8: e54059.

37. de Assis AM, Rieger DK, Longoni A, Battu C, Raymundi S, da Rocha RF. High fat and highly thermolyzed fat diets promote insulin resistance and increase DNA damage in rats. Exp Biol Med (Maywood). 2009;234: 1296–1304.

38. Włodarczyk M, Jabłonowska-Lietz B, Olejarz W, Nowicka G. Anthropometric and dietary factors as predictors of DNA damage in obese women. Nutrients. 2018;10. doi:10.3390/nu10050578

39. Kompella P, Wang G, Durrett RE, Lai Y, Marin C, Liu Y. Obesity increases genomic instability at DNA repeat-mediated endogenous mutation hotspots. Nat Commun. 2024;15: 6213.

40. Williams EG, Pfister N, Roy S, Statzer C, Haverty J, Ingels J. Multiomic profiling of the liver across diets and age in a diverse mouse population. Cell Syst. 2022;13: 43–57.e6.

41. Muyas F, Sauer CM, Valle-Inclán JE, Li R, Rahbari R, Mitchell TJ. De novo detection of somatic mutations in high-throughput single-cell profiling data sets. Nat Biotechnol. 2023. doi:10.1038/s41587-023-01863-z

42. Tabula Sapiens Consortium*, Jones RC, Karkanias J, Krasnow MA, Pisco AO, Quake SR. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. Science. 2022;376: eabl4896.

43. Gabitto MI, Travaglini KJ, Rachleff VM, Kaplan ES, Long B, Ariza J. Integrated multimodal cell atlas of Alzheimer's disease. Nat Neurosci. 2024;27: 2366–2383.

44. De Jager PL, Ma Y, McCabe C, Xu J, Vardarajan BN, Felsky D. A multi-omic atlas of the human frontal cortex for aging and Alzheimer's disease research. Sci Data. 2018;5: 180142.

45. Perampalam P, Hassan HM, Lilly GE, Passos DT, Torchia J, Kiser PK. Disrupting the DREAM transcriptional repressor complex induces apolipoprotein overexpression and systemic amyloidosis in mice. J Clin Invest. 2021;131. doi:10.1172/JCI140903

46. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV. Signatures of mutational processes in human cancer. Nature. 2013;500: 415–421.

47. Guo Q, Lakatos E, Bakir IA, Curtius K, Graham TA, Mustonen V. The mutational signatures of formalin fixation on the human genome. Nat Commun. 2022;13: 4487.

48. Alexandrov LB, Jones PH, Wedge DC, Sale JE, Campbell PJ, Nik-Zainal S. Clock-like mutational processes in human somatic cells. Nat Genet. 2015;47: 1402–1407.

49. Everall A, Tapinos A, Hawari A, Cornish A, Sud A, Chubb D. Comprehensive repertoire of the chromosomal alteration and mutational signatures across 16 cancer types from 10,983 cancer patients. bioRxiv. 2023. p. 2023.06.07.23290970. doi:10.1101/2023.06.07.23290970

50. Krokan HE, Bjørås M. Base excision repair. Cold Spring Harb Perspect Biol. 2013;5: a012583.

51. Nicolas E, Golemis EA, Arora S. POLD1: Central mediator of DNA replication and repair, and implication in cancer and other pathologies. Gene. 2016;590: 128–141.

52. Clauson C, Schärer OD, Niedernhofer L. Advances in understanding the complex mechanisms of DNA interstrand cross-link repair. Cold Spring Harb Perspect Biol. 2013;5: a012732.

53. Davis AJ, Chen DJ. DNA double strand break repair via non-homologous end-joining. Transl Cancer Res. 2013;2: 130–143.

54. Powers KT, Washington MT. Eukaryotic translesion synthesis: Choosing the right tool for the job. DNA Repair (Amst). 2018;71: 127–134.

55. Mao P, Wyrick JJ, Roberts SA, Smerdon MJ. UV-induced DNA damage and Mutagenesis in chromatin. Photochem Photobiol. 2017;93: 216–228.

56. Sekiguchi M, Tsuzuki T. Oxidative nucleotide damage: consequences and prevention. Oncogene. 2002;21: 8895–8904.

57. Viel A, Bruselles A, Meccia E, Fornasarig M, Quaia M, Canzonieri V. A specific mutational signature associated with DNA 8-oxoguanine persistence in MUTYH-defective colorectal cancer. EBioMedicine. 2017;20: 39–49.

58. Zhang H. Mechanisms of mutagenesis induced by DNA lesions: multiple factors affect mutations in translesion DNA synthesis. Crit Rev Biochem Mol Biol. 2020;55: 219–251.

59. Hulbert AJ, Pamplona R, Buffenstein R, Buttemer WA. Life and death: metabolic rate, membrane composition, and life span of animals. Physiol Rev. 2007;87: 1175–1213.

60. Speakman J. Body size, energy metabolism and lifespan. J Exp Biol. 2005;208: 1717–1730.

61. MacRae SL, Croken MM, Calder RB, Aliper A, Milholland B, White RR. DNA repair in species with extreme lifespan differences. Aging (Albany NY). 2015;7: 1171–1184.

62. Firsanov D, Zacher M, Tian X, Sformo TL, Zhao Y, Tombline G. DNA repair and anti-cancer mechanisms in the long-lived bowhead whale. bioRxivorg. 2024; 2023.05.07.539748.

63. Tall AR, Fuster JJ. Clonal hematopoiesis in cardiovascular disease and therapeutic implications. Nat Cardiovasc Res. 2022;1: 116–124.

64. Jaiswal S, Natarajan P, Silver AJ, Gibson CJ, Bick AG, Shvartz E. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. N Engl J Med. 2017;377: 111–121.

65. Zhao Y, Simon M, Seluanov A, Gorbunova V. DNA damage and repair in age-related inflammation. Nat Rev Immunol. 2023;23: 75–89.

66. Lovell MA, Markesbery WR. Oxidative DNA damage in mild cognitive impairment and late-stage Alzheimer's disease. Nucleic Acids Res. 2007;35: 7497–7504.

67. Asada-Utsugi M, Uemura K, Ayaki T, T Uemura M, Minamiyama S, Hikiami R. Failure of DNA double-strand break repair by tau mediates Alzheimer's disease pathology in vitro. Commun Biol. 2022;5: 358.

68. Ogawa Y, Nonaka Y, Goto T, Ohnishi E, Hiramatsu T, Kii I. Development of a novel selective

inhibitor of the Down syndrome-related kinase Dyrk1A. Nat Commun. 2010;1: 86.

69. Seifert A, Allan LA, Clarke PR. DYRK1A phosphorylates caspase 9 at an inhibitory site and is potently inhibited in human cells by harmine. FEBS J. 2008;275: 6268–6280.

70. Réus GZ, Stringari RB, de Souza B, Petronilho F, Dal-Pizzol F, Hallak JE. Harmine and imipramine promote antioxidant activities in prefrontal cortex and hippocampus. Oxid Med Cell Longev. 2010;3: 325–331.

71. Caruso A, Caira V, El-Kashef H, Saturnino C. The potential of indole alkaloids in bone health and osteoporosis management. Appl Sci (Basel). 2024;14: 8124.

72. Waki H, Park KW, Mitro N, Pei L, Damoiseaux R, Wilpitz DC. The small molecule harmine is an antidiabetic cell-type-specific regulator of PPARγ expression. Cell Metab. 2007;5: 357–370.

73. Morsy MHE, Nabil ZI, Darwish ST, Al-Eisa RA, Mehana AE. Anti-Diabetic and Anti-Adipogenic Effect of Harmine in High-Fat-Diet-Induced Diabetes in Mice. Life. 2023;13. doi:10.3390/life13081693

74. Branca C, Shaw DM, Belfiore R, Gokhale V, Shaw AY, Foley C. Dyrk1 inhibition improves Alzheimer's disease-like pathology. Aging Cell. 2017;16: 1146–1154.

75. de la Torre R, de Sola S, Hernandez G, Farré M, Pujol J, Rodriguez J. Safety and efficacy of cognitive training plus epigallocatechin-3-gallate in young adults with Down's syndrome (TESDAD): a double-blind, randomised, placebo-controlled, phase 2 trial. Lancet Neurol. 2016;15: 801–810.

76. Stotani S, Giordanetto F, Medda F. DYRK1A inhibition as potential treatment for Alzheimer's disease. Future Med Chem. 2016;8: 681–696.

77. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19: 15.

78. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102: 15545–15550.

79. Lause J, Berens P, Kobak D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome Biol. 2021;22: 258.

80. Li Y, Dou Y, Da Veiga Leprevost F, Geffen Y, Calinawan AP, Aguet F. Proteogenomic data and resources for pan-cancer analysis. Cancer Cell. 2023;41: 1397–1406.

81. Gabitto MI, Travaglini KJ, Rachleff VM, Kaplan ES, Long B, Ariza J. Integrated multimodal cell atlas of Alzheimer's disease. Res Sq. 2023; rs.3.rs–2921860.

82. Dong X, Zhang L, Milholland B, Lee M, Maslov AY, Wang T. Accurate identification of single-nucleotide variants in whole-genome-amplified single cells. Nat Methods. 2017;14: 491–493.

83. Ren P, Dong X, Vijg J. Age-related somatic mutation burden in human tissues. Front Aging. 2022;3: 1018119.

84. Lawson ARJ, Abascal F, Nicola PA, Lensing SV, Roberts AL, Kalantzis G. Somatic mutation and selection at epidemiological scale. Genetic and Genomic Medicine. medRxiv; 2024. Available: https://www.medrxiv.org/content/10.1101/2024.10.30.24316422v2.full.pdf

85. Ganz J, Luquette LJ, Bizzotto S, Miller MB, Zhou Z, Bohrson CL. Contrasting somatic mutation patterns in aging human neurons and oligodendrocytes. Cell. 2024;187: 1955–1970.e23.

86. Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing SV, Ellis P. Somatic mutation landscapes at single-molecule resolution. Nature. 2021;593: 405–410.

87. Heilbrun EE, Merav M, Adar S. Exons and introns exhibit transcriptional strand asymmetry of dinucleotide distribution, damage formation and DNA repair. NAR Genom Bioinform. 2021;3:

lqab020.

88. Frigola J, Sabarinathan R, Mularoni L, Muiños F, Gonzalez-Perez A, López-Bigas N. Reduced mutation rate in exons due to differential mismatch repair. Nat Genet. 2017;49: 1684–1692.

89. pylr2: Linear Regression Type II Models. Github; Available: https://github.com/OceanOptics/pylr2

90. de Magalhães JP, Costa J, Church GM. An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. J Gerontol A Biol Sci Med Sci. 2007;62: 149–160.

91. Davidson-Pilon C. lifelines: survival analysis in Python. J Open Source Softw. 2019;4: 1317.

92. Montine TJ, Phelps CH, Beach TG, Bigio EH, Cairns NJ, Dickson DW. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease: a practical approach. Acta Neuropathol. 2012;123: 1–11.

93. Cheng Y, Nandi SP, Culibrk L, Kristin A, Stuewe I, Al-Azzam S. Improved mutation detection in duplex sequencing data with sample-specific error profiles. bioRxivorg. 2025. p. 2025.07.13.664565. doi:10.1101/2025.07.13.664565

94. Cook RD. Detection of influential observation in linear regression. Technometrics. 1977;19: 15–18.

95. Cao H, Deng B, Song T, Lian J, Xia L, Chu X. Genome-wide profiles of DNA damage represent highly accurate predictors of mammalian age. Aging Cell. 2024;23: e14122.

96. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28: 27–30.

97. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25: 25–29.

Chapter 3 contains unpublished material coauthored with Koch, Zane; Nandi, Shuvro P.; Licon, Kate; Bujarrabal-Dueso, Arturo; Meyer, David H.; Saeed, Safa; Perampalam, Pirunthan;

Acknowledgements

Dick, Frederick A.; Schumacher, Björn; Alexandrov, Ludmil B.; Ideker, Trey. The dissertation author was the primary author of this chapter.