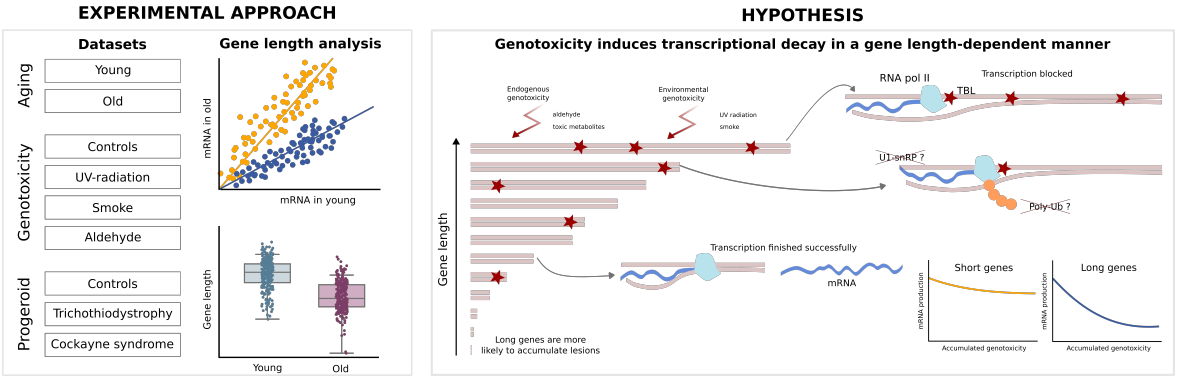
Age- or lifestyle-induced accumulation of genotoxicity is associated with a generalized shutdown of long gene transcription

Olga Ibañez-Solé[1], Ander Izeta[1,2*]

1 Biodonostia Health Research Institute, Tissue Engineering Group, Donostia-San Sebastián, Spain
2 Tecnun-University of Navarra, Donostia-San Sebastián, Spain.

* ander.izeta@biodonostia.org



Abstract

A causative role for DNA damage as a molecular driver of aging has long been advocated. Transcription-blocking lesions (TBLs) accumulate with age in a stochastic manner. Thus, gene expression data might reflect the gene length-dependent accumulation of TBLs. Here we present an analysis of gene expression as a function of gene length in several independent single-cell RNA sequencing datasets of mouse and human aging. We found a pervasive age-associated downregulation of long gene expression, which is seen across species, datasets, sexes, tissues and cell types. Furthermore, long gene downregulation was also observed in premature aging models such as UV-radiation and smoke exposure, and in gene expression data from progeroid diseases Cockayne syndrome and trichothiodystrophy. Finally, we analyzed the length of differentially expressed genes associated to age in both mice and humans. Downregulated genes were significantly longer than upregulated genes. These data highlight a previously undetected hallmark of cellular aging and provide strong support for age-associated accumulation of genotoxic damage inducing a generalized shutdown of RNA polymerase II-mediated long gene transcription.

Introduction

DNA damage has long been proposed as a primary molecular driver of aging [1, 2]. Aging has also been associated with a series of transcriptional changes, most of which are highly tissue- and cell type-specific [3]. Even though the search for a global aging signature has been the goal of much research [4, 5, 6, 7], meta-analyses have shown that very few genes are consistently up- or downregulated with aging across different tissues [8]. It appears that, at the mRNA level, aging signatures are not defined by the overexpression of particular sets of genes – in fact, the differences between the transcriptome of middle-aged and young individuals are bigger than those between young and old individuals, at least in some human tissues [9]– but rather, an overall decay in transcription [10].

Genetic material is constantly challenged throughout the lifespan of the organism, both by endogenous and environmental genotoxins. Some of this damage happens in the form of transcription-blocking lesions (TBLs), which impede transcriptional elongation [11]. Accumulation of TBLs provokes a genome-wide shutdown of transcription which also affects undamaged genes through poorly understood mechanisms, that may be related to RNA polymerase II (RNAP II) ubiquitylation and degradation [12, 13]. Assuming a constant TBL incidence, meaning that any base pair in the genome has a similar probability of suffering damage that results in a lesion, a greater accumulation of TBLs is to be expected in longer genes. As a matter of fact, a gene length-dependent accumulation of other forms of genetic damage, like somatic mutations, has already been reported in conditions like Alzheimer's disease [14]. Hence, TBLs, just like somatic mutations, are expected to accumulate with aging, and their accumulation is expected to be dependent on gene length. However, unlike somatic mutations, TBLs have a strong and direct impact on mRNA production, and their gene length-dependent effects are likely to be measurable from RNA sequencing data of aged tissues, which make single-cell RNA sequencing (scRNA-seq) atlases and datasets of aging an excellent opportunity to characterize them at the cell type level over a wide range of tissues.

So far, a potential relationship between age-related transcriptional changes and gene length has received relatively little attention. A recent analysis of the transcript length of 307 genes related to aging (as extracted from the *GenAge* database) found longer transcript lengths in these genes as compared to the rest of the protein-coding genes [15]. However, when they studied aging gene-expression signatures from a human, mouse and rat meta-analysis, they found no significance regarding transcript length in overexpressed and underexpressed genes, the only exception being the brain (which downregulated long genes). Of interest, a previous analysis of gene expression profiles in the liver of mice deficient in the DNA excision-repair gene *Ercc1*, which present features of accelerated aging, had found specific downregulation of long genes [16]. Similar findings were reported by the authors in naturally aged rat liver and human hippocampus, indicating that it could reflect a more generalized phenomenon. Here we aimed to extend these early observations, which were based on bulk microarray and RNA sequencing data, to the existing aging datasets based on scRNA-seq technology. We also extended our gene length analyses to mouse and human datasets of lifestyle-induced genotoxic exposure (UV, smoke) and progeroid syndromes (Cockayne Syndrome and trichothiodystrophy).

Results

Age-associated shutdown of transcription preferentially affects long genes

In order to test if gene expression at the single-cell level is conserved with aging, we first analyzed 11 organs of the landmark *Tabula Muris Senis (TMS)* dataset of mouse aging [17], on the basis of having enough experimental replicates and single cells for statistically significant analyses. Thus, we selected male animals of both young (3-month) and old (24-month) age (Figure 1). Plotting the average gene expression of aged tissues against their young counterparts yielded scatter plots where data presented a high linear correlation between both average expression vectors (Figure 1a). However, we observed that a large number of genes lied below the $y = x$ line, meaning that their mean expression was lower in old mice. This was most evident in brain, heart, liver, lung, muscle, pancreas and skin. Having established that there is an age-related decline in mRNA production, we explored the gene-length dependence of such decline. To this end, we split the whole transcriptome into four equally sized bins according to gene length and fitted a multiple linear regression model considering the interaction effect between average expression in young and the categorical variable representing the gene-length quartile. We found that the slope of the straight line that fits the gene expression data decreases with gene length, which confirms that the decay in mRNA production is strongly dependent on gene length. We graphically show this difference for the two most extreme quartiles (25% shortest and the 25% longest genes) in Figure 1b; gene lengths and p values for all comparisons are shown in Supplementary Tables S1 and S2). The differences in gene lengths were statistically significant in all analyzed organs.

This effect was also detected in independent scRNA-seq datasets obtained from mouse lung, kidney, spleen and skin [18, 19, 20, 21], although there were relevant experimental differences among datasets (Supplementary Figure S1). Importantly, downregulation of longer genes was also evident in single-cell data of human lung, pancreas and skin [22, 23, 24, 25] (Supplementary Figure S1). Similarly, the effect was also detectable in *TMS* female animals (Supplementary Figure S2). These results suggested a generalized downregulation of long gene expression associated with age, which is seen across tissues, sexes and species, and in data extracted from several independent scRNA-seq datasets.

Differentially expressed genes between young and old individuals show a preferential bias for the downregulation of long genes

A number of genes change their expression in the same direction during aging in several tissues, and the search for differentially expressed genes (DEGs) may thus provide a molecular signature of aging [26]. We next analyzed if DEGs between young and old animals from the *TMS* dataset showed a preferential bias for the downregulation of long genes. Indeed, that was the case, since DEGs between young (3-month) and old (24-month) mice showed a statistically significant bias for the downregulation of long genes for all tissues and comparisons based on a Wilcoxon-Mann-Whitney test (Figure 2, p-values are provided in the Supplementary Table S3). Once more, this effect was not specific of the *TMS* dataset, since it was also detected in independent scRNA-seq datasets obtained from mouse lung, kidney, spleen and skin and human lung, pancreas and skin (Supplementary Figure S3). Finally, the effect was also detectable in *TMS* female animals (Supplementary Figure S4). Despite the fact that inter-individual and inter-tissue differences were apparent in some cases, these data confirmed that long genes were differentially affected by the age-associated shutdown of transcription.

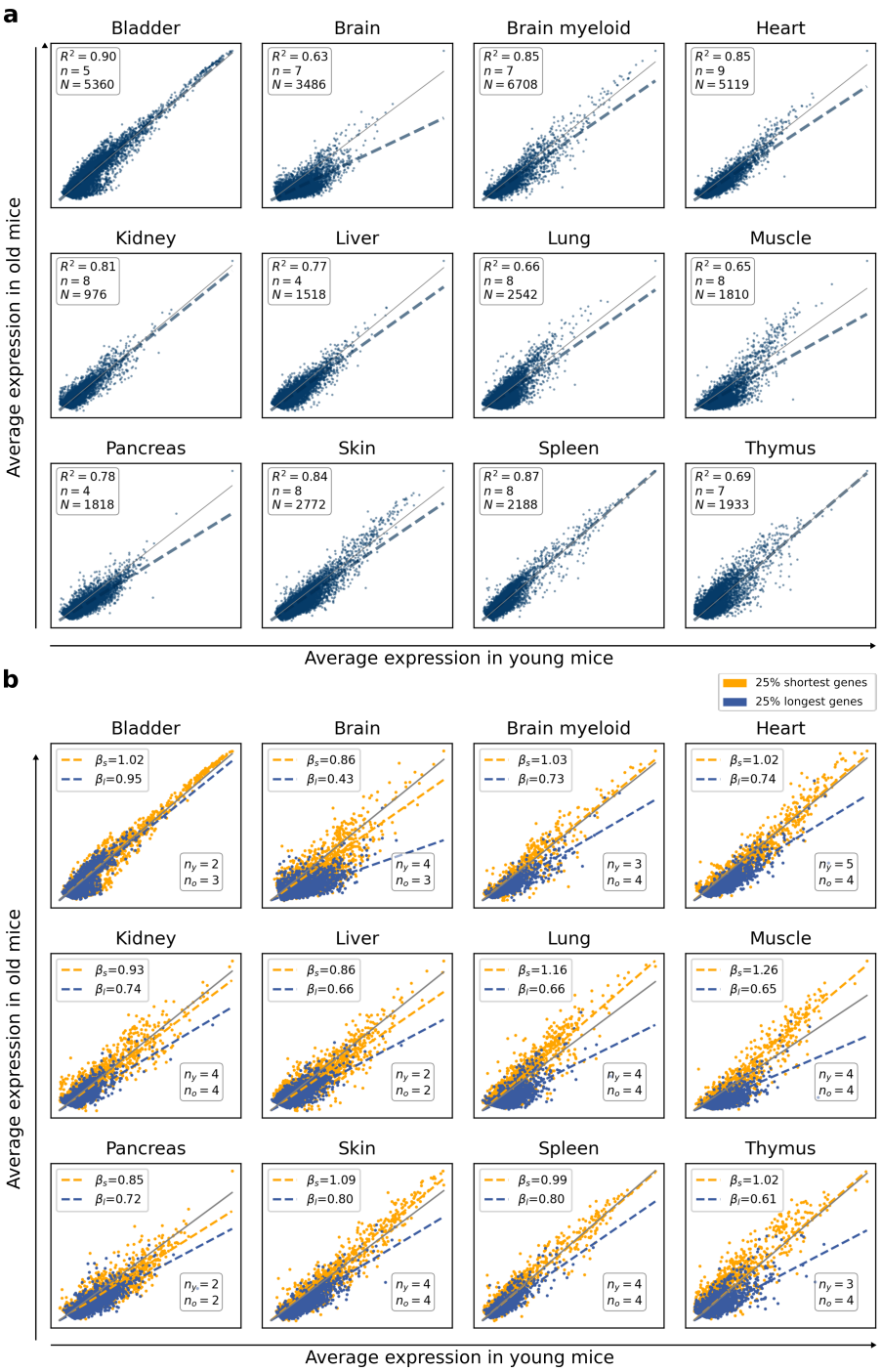


Figure 1. A generalized, age-associated shutdown of long gene transcription. **a**, Gene expression is highly conserved but shows a detectable decay with aging. Scatter plots showing the average gene expression in 24-month old mice against average gene expression in 3 month-old mice in 11 tissues from the *TMS FACS* and the *TMS droplet* datasets [17]. Each dot represents a gene. N: number of single cells; n: number of biological replicates. R^2: coefficient of determination. Thre grey line represents $y = x$. **b**, A generalized shutdown of transcription is apparent in long genes. The scatter plots show the average gene expression of the 25% shortest (yellow) and the 25% longest (blue) genes in 24 month-old versus in 3 month-old mice. β_s and β_l represent the slopes of the straight lines that best fit the data points corresponding to *short* and *long* genes, respectively. The number of young (n_y) and old (n_o) biological replicates are shown.

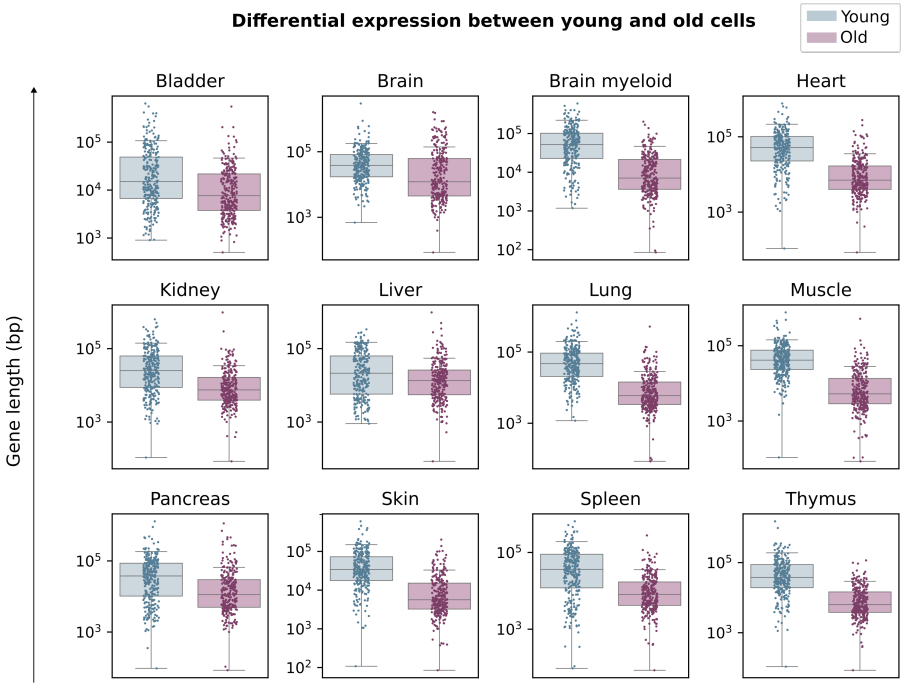


Figure 2. Differentially expressed genes between young and old animals show a preferential bias for the downregulation of long genes. Top 300 DEGs between young and old cells in 12 aging datasets from the *Tabula Muris Senis*. The 300 differentially expressed genes between 3 months old and 24 months old male mice were obtained using the Wilcoxon method. The difference between young and old DEG length is significant in all tissues (p-value < 0.001), see Supplementary table S3

The age-associated decrease in the expression of long genes is not cell type-specific

Since many aging signatures are cell type-specific, a relevant open question was if the age-associated downregulation of long genes might be restricted to a particular cell type that is abundantly and ubiquitously located across tissues, such as fibroblasts or endothelial cells. To answer this question, we selected the four existing *TMS* heart datasets and analyzed the gene length of expressed genes (Figure 3). As expected, shorter genes were overexpressed in old mice as compared to young mice in all four datasets (Figure 3a). Compartmentalization of the analyses onto the 11 single-cell types detected in at least two datasets showed that young animals expressed longer genes in all cell types analyzed, including tissue-specific cells such as cardiomyocytes and infiltrating cell types such as B and T lymphocytes (Figure 3b). Therefore, a pervasive downregulation of long genes was detectable across aged tissues and cell types.

Genotoxic UV exposure of young mouse skin mimics age-associated decrease in the expression of long genes

Ultraviolet (UV) radiation of skin exposed to sunlight produces accumulation of DNA damage and photoaging [27, 28]. Notably, UV-induced photolesions – mainly cyclobutane pyrimidine dimers (CPDs) and pyrimidine-(6-4)-pyrimidone photoproducts (6-4PPs)– trigger a general shutdown of transcription and are mainly repaired by the Nucleotide Excision Repair (NER) pathways [13] . The vitamin D

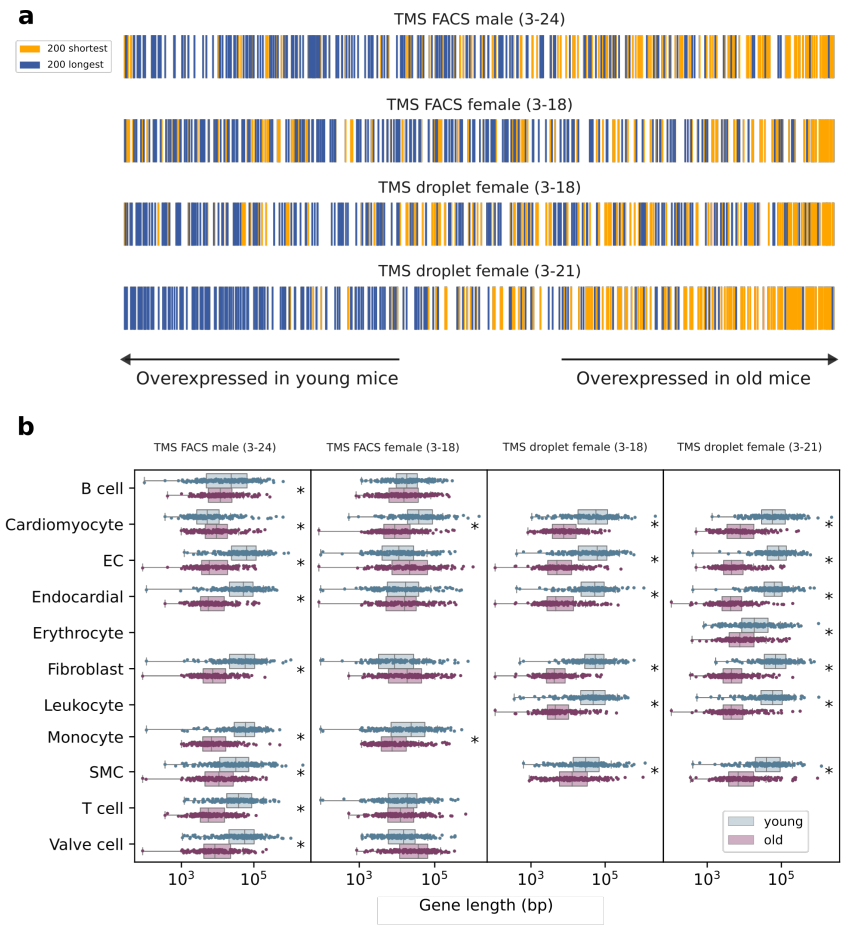


Figure 3. The age-associated decrease in the expression of long genes is not cell type-specific. a, Genes ranked according to their age-related difference in average gene expression. Genes are shown sorted according to their difference in mean expression between old and young cells. The positions of the top 200 shortest (yellow) and the top 200 longest (blue) genes are shown. **b**, Genes differentially expressed between old and young cells have significantly different gene lengths. Gene length of the 200 most differentially expressed genes (DEGs) between young and old cells within each cell type. EC, endothelial cell. SMC, smooth muscle cell. Significant differences (Wilcoxon-Mann-Whitney test, $p - value < 0.01$) are marked with an asterisk (*).

system provides a local adaptive response to UV radiation, reducing DNA damage, inflammation and photocarcinogenesis [29]. To test if genotoxic damage to DNA (a premature aging model) also affected the transcription of long genes, we analyzed a single-cell RNAseq dataset of young (five to six-week-old) mouse skin irradiated with UVB or normal light [30]. One of the UV-irradiated groups was injected with vitamin D (Figure 4). A Uniform manifold approximation and projection (UMAP) plot of the merged datasets of mice skin shows the 11 cell types detected in this experiment using unsupervised cell clustering (Figure 4a). Of note, long gene expression decreased in UV-radiated skin as compared to both healthy and vitamin D-treated groups (Figure 4b-c). An analysis of the length of the top 300 DEGs computed between the three conditions (the genes differentially expressed in each of the conditions against the remaining two) further demonstrated that longer genes were differentially affected by UVB exposure (Figure 4d-e). Finally, this effect was detected in all skin cell types, although not all long gene transcriptional phenotypes were rescued by vitamin D injection (Figure 4f). These results strongly suggested that environmental genotoxic damage by UV-radiation may induce a generalized shutdown of long gene transcription in young animals, which may be partially

reverted by vitamin D injection.

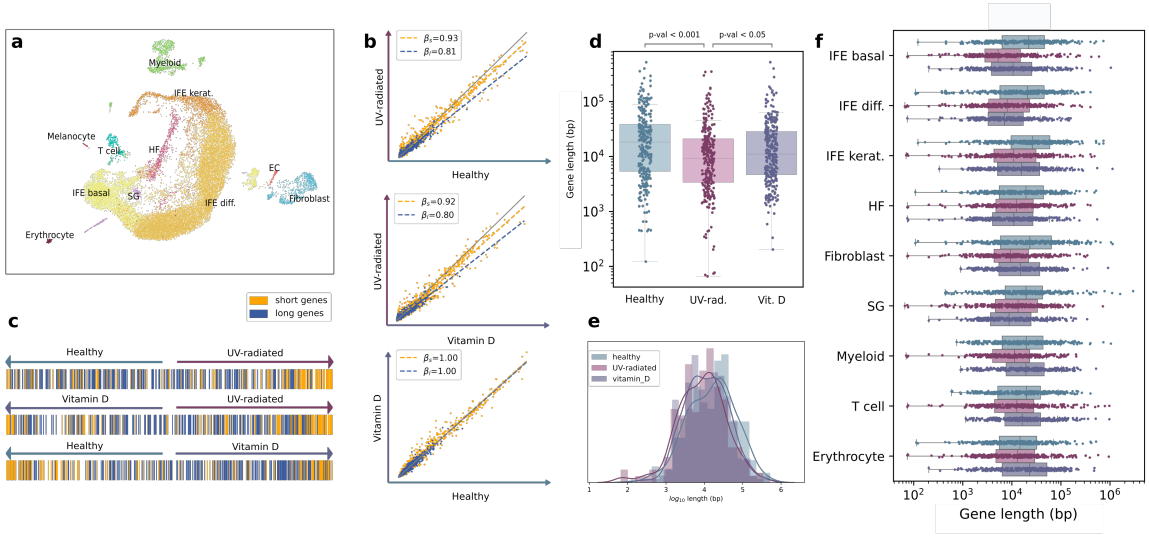


Figure 4. Genotoxic UV exposure of young mouse skin decreases the expression of long genes. a, Uniform manifold approximation and projection (UMAP) plot showing 11 cell types in the murine skin dataset [30]. The samples corresponding to the three conditions (healthy, UV-radiated and UV-radiated with a vitamin D treatment) were merged into a single dataset. Diff., differentiated. EC, endothelial cell. HF, hair follicle. IFE, interfollicular epidermis. Kerat., keratinocytes. SG, sebaceous gland. **b**, Long-gene expression decreases in UV-radiated skin, but not in vitamin D-treated skin. Scatter plots showing the mean expression in every pair of conditions: UV-radiated *vs* healthy skin (top), UV-radiated *vs* vitamin D-treated skin (middle) and vitamin D-treated *vs* healthy skin (bottom). β_s and β_l correspond to the slopes of the multiple linear regression models with interaction fitted on the 1st and 4th quartiles (top 25% shortest and top 25% shortest genes). **c**, Shortest genes are overexpressed in UV-radiated skin. Position of the top 200 shortest and top 200 longest genes, in the differential expression ranking. Genes are shown ranked according to their difference in mean expression between every pair of conditions. Genes are colored according to their length: top 200 shortest (yellow) and top 200 longest (blue). **d**, Length of the genes differentially expressed in UV-radiated skin cells. Top 300 DEGs are computed between the three conditions (those differentially expressed in each of the conditions against the remaining two). The distributions of log_{10} gene length (bp) is shown. The p-values were obtained in a Tukey post-hoc test after ANOVA. **e**, Log-transformed gene lengths for the DEGs associated with the three conditions are normally distributed. A histogram and a density plot are shown for each condition. The three distributions are normal (Lilliefors normality test, p-value $>$ 0.05). **f**, DEGs associated with UV-radiated skin cell types are significantly shorter. The DEGs were computed between the three conditions for each cell type separately.

Smoke exposure of human airways mimics age-associated decrease in the expression of long genes

Chronological age of never-smokers does increase the frequency of mutations in bronchial epithelial cells at a rate of 28 mutations per cell per year. Mutation frequency in cells from smokers increased at a rate of 91 mutations per cell per year, i.e. 3.25X higher [31]. In addition to somatic mutations, exposure to smoke from organic matter is known to provoke TBLs [11], due to benzo[a]pyrene diol epoxide (BPDE) reacting with guanines to form bulky DNA adducts [13]. To test if the lifestyle of smokers affected specifically the expression of long genes in airway epithelial cells, we analyzed a scRNA-seq dataset [32] of human trachea of never-smokers and heavy smokers (subjects who had been

smoking for >20 years) of a similar age range (Figure 5). A UMAP plot of the merged datasets of both never-smokers and heavy smokers detected 13 cell types in human trachea (Figure 5a). As expected by their increased accumulated genotoxicity, long gene expression significantly decreased in heavy smokers as compared to never-smokers (Figure 5b-c, p-values in Supplementary Table S4). An analysis of the length of the top 300 DEGs computed between both groups further demonstrated that longer genes were differentially affected by smoke exposure (Figure 5d-e). Finally, this effect was not cell-specific since it was detected in all tracheal cell types (Figure 5f). These results confirmed that environmental genotoxic damage induces a generalized shutdown of long gene transcription.

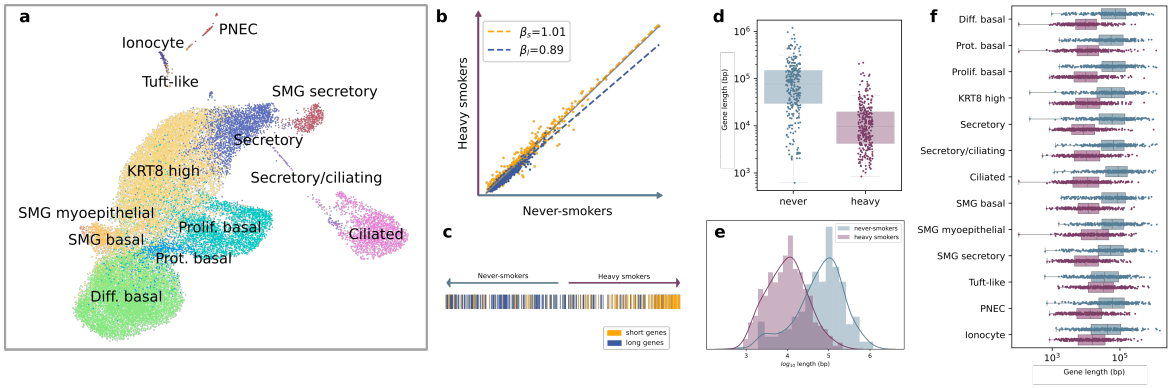


Figure 5. Smoke exposure of human airway epithelial cells mimics age-associated decrease in the expression of long genes. a, UMAP showing the 13 detected cell types in the human trachea dataset. The samples corresponding to the two conditions (never-smokers and heavy smokers) were merged into a single dataset. Diff, differentiated. KRT8, Keratin 8. PNEC, pulmonary neuroendocrine cells. Prolif., proliferating. Prot., proteasomal. SMG, submandibular salivary glands. **b**, Long-gene expression in decreased in heavy smokers. The scatter plot shows the average gene expression in heavy smokers *vs* average gene expression in never-smokers. β_s and β_l correspond to the slopes of the linear regression models fitted on the 1st and 4th quartiles (top 25% shortest and top 25% shortest genes). **c**, Shortest genes are overexpressed in airway cells from heavy smokers. Position of the top 200 shortest (yellow) and top 200 longest (blue) genes in the differential expression ranking. **d**, The length of the genes differentially expressed in airway cells from heavy smokers *vs* never-smokers. 300 DEGs are computed between the two conditions. The distributions of log_{10} gene length (bp) is shown. The p-values were obtained in a Mann Whitney U test. **e**, Log-transformed gene lengths for the DEGs associated with never smokers are not normally distributed. A histogram and a density plot are shown for each condition. Only the distribution of the DEG lengths for heavy smokers passed the Lilliefors normality test. **f**, DEGs associated with heavy smoker airway cell types are significantly shorter. The DEGs were computed between never-smokers and heavy smokers for each cell type separately.

Transcriptional stress in progeroid diseases Cockayne Syndrome and trichothiodystrophy results in a decrease in the expression of long genes

A number of progeroid diseases are caused by mutations functionally linked to genome maintenance and DNA damage repair [33]. Of particular interest to this work, a subset of defects in repair genes impair transcription-coupled nucleotide excision repair (TC-NER), i.e. TBLs remain unrepaired, causing RNAPII stalling and ultimately syndromic features such as Cockayne Syndrome, xeroderma pigmentosum, and trichothiodystrophy [11]. Of interest, increased cutaneous photosensitivity is one of the clinical features of patients suffering from these conditions, and is caused by deficiencies in genes

coding for components of the TC-NER.

Endogenous formaldehyde is abundant in the body, causing DNA crosslinks, oxidative stress and potentially contributing to the onset of Fanconi Anemia and other syndromes [34]. On the other hand, Cockayne Syndrome is caused by loss of the Cockayne Syndrome A (CSA) or CSB proteins. Of note, double knock-out mice deficient in both formaldehyde clearance ($Adh5^{-/-}$) and CSB protein ($Csb^{m/m}$) develop transcriptional stress in a subset of kidney cells and features consistent with human Cockayne Syndrome [35]. To test if kidney cells of these animals undergoing formaldehyde-driven transcriptional stress specifically decreased transcription of long genes, we analyzed single-cells of three knockout mice – *ADH5KO* (deficient in formaldehyde clearance), *CSBKO* (Cockayne Syndrome group B knock-out, also known as *Ercc6*), and *DKO* ($Adh5^{-/-}Csb^{m/m}$ double knock-out) – against those of wild type (*WT*) mice (Figure 6). Interestingly, specific downregulation of long genes was already detected in *ADH5KO* and *CSBKO* single mutants. Both mutations seemed to synergize causing further downregulation of long genes in the *DKO* animals as compared to *WT* mice (Figure 6A-B, *p*-values in Supplementary Table S4). An analysis of the length of the top 300 DEGs computed between *WT* and *ADH5KO*, *WT* and *CSBKO*, and *WT* and *DKO* groups further demonstrated that longer genes were differentially affected by formaldehyde-driven transcriptional stress (Figure 6C).

Encouraged by these results, we analyzed a microarray dataset of human mesenchymal stromal cells (MSCs) derived from a Cockayne Syndrome patient bearing a *CSB/ERCC6* mutation, which are known to present marked changes in their transcriptome upon UV-radiation [36]. In fact, skin fibroblasts from this patient were first reprogrammed to generate induced pluripotent stem cells, which were then gene-corrected with CRISPR-Cas9, and differentiated to MSCs. Thus, the available data included UV-radiated MSCs *vs* MSCs in normal conditions in both mutant ($ERCC^{mut}$) and gene-corrected ($ERCC^{GC}$) backgrounds (Figure 7). As expected, UV-radiation on $ERCC^{mut}$ cells induced a decrease in long gene expression as compared to normal conditions in both mutant and gene-corrected ($ERCC^{GC}$) cells (Figure 7a). Plotting of the gene lengths of the top 300 DEGs between cells with and without UV-radiation exposure in both mutant and gene corrected backgrounds, further demonstrated a bias for long gene downregulation (Figure 7b-c). This was due to the combined effect of UV-radiation and *CSB/ERCC6* mutation, since comparisons between mutant ($ERCC^{mut}$) and gene corrected ($ERCC^{GC}$) cells in normal conditions (control) and after UV-radiation exposure demonstrated that GC-cells were unaffected in control conditions (Figure 7d). An analysis of the length of the 300 most differentially expressed genes between mutant and gene-corrected cells further illustrated this point (Figure 7e-f). Overall, these results demonstrated that transcriptional stress provided by aldehyde and UV-radiation in Cockayne Syndrome preferentially affected the transcription of long genes.

Finally, we tested if long gene transcription was also affected in a second progeroid syndrome, trichothiodystrophy (TTD). To this end, we analyzed the length of the DEGs obtained by Lombardi et al. [37] between a cancer-free photosensitive trichothiodystrophy (PS-TTD) patient carrying a mutation in the *ERCC2* gene and her healthy mother, both in basal conditions and upon UV-radiation. Selecting the genes that were significantly (p-value ≤ 0.05) over- or underexpressed in PS-TTD and with a substantial effect size (logFC ≥ 2 in either direction), we observed that the DEGs associated with PS-TTD were significantly shorter upon UV-radiation (Figure 7g). These results suggested that other progeroid syndromes may present a similar phenotype of reduced long gene transcription.

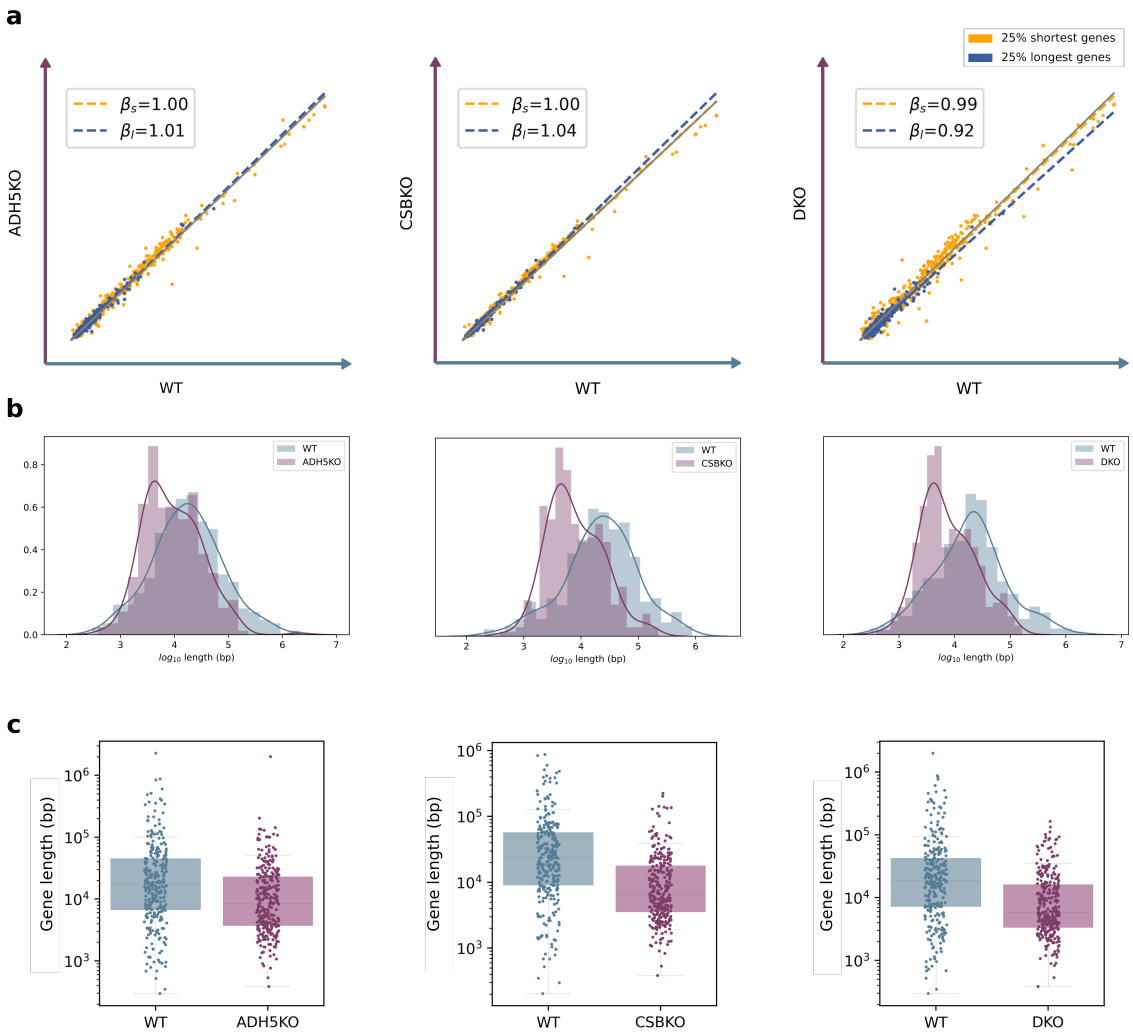


Figure 6. A mouse model of Cockayne Syndrome mimics age-associated decrease in the expression of long genes in the kidney. a, Correlation between global average gene expression in each of the knock-outs against the wild type mice. The average gene expression in three knockout mice – *ADH5KO* (deficient in formaldehyde clearance) *CSBKO* (Cockayne Syndrome group B knock-out, also known as *Ercc6*) and *DKO* ($Adh5^{-/-}Csb^{m/m}$ double knock-out) – against wild type mice. Each data point represent a gene. β_s and β_l represent the slopes of the straight lines that best fit the data points corresponding to the 25% shortest (yellow) and 25% longest (blue) genes, respectively. **b-c**, Distribution of gene lengths in the genes differentially overexpressed in each of the knock-outs *vs* the wild type mice. The log-transformed gene length of the 300 most differentially expressed genes between each of the knock outs and the wild type mice are shown in a density plot over a histogram (**b**) and a stripplots over boxplots (**c**).

Published aging signatures are influenced by gene length-dependent transcriptional decay

A number of aging-related transcriptional signatures have been proposed for both mice and humans. A recent study identified a set of mouse *global aging genes (GAGs)* [26], defined as genes whose expression varies substantially with age in most (>50%) of the tissue-cell types across several tissues of the *TMS* dataset. They found that GAGs exhibited a strong bimodality, i.e., that they were either upregulated or downregulated with aging in most tissues. However, to our knowledge no study of gene length has

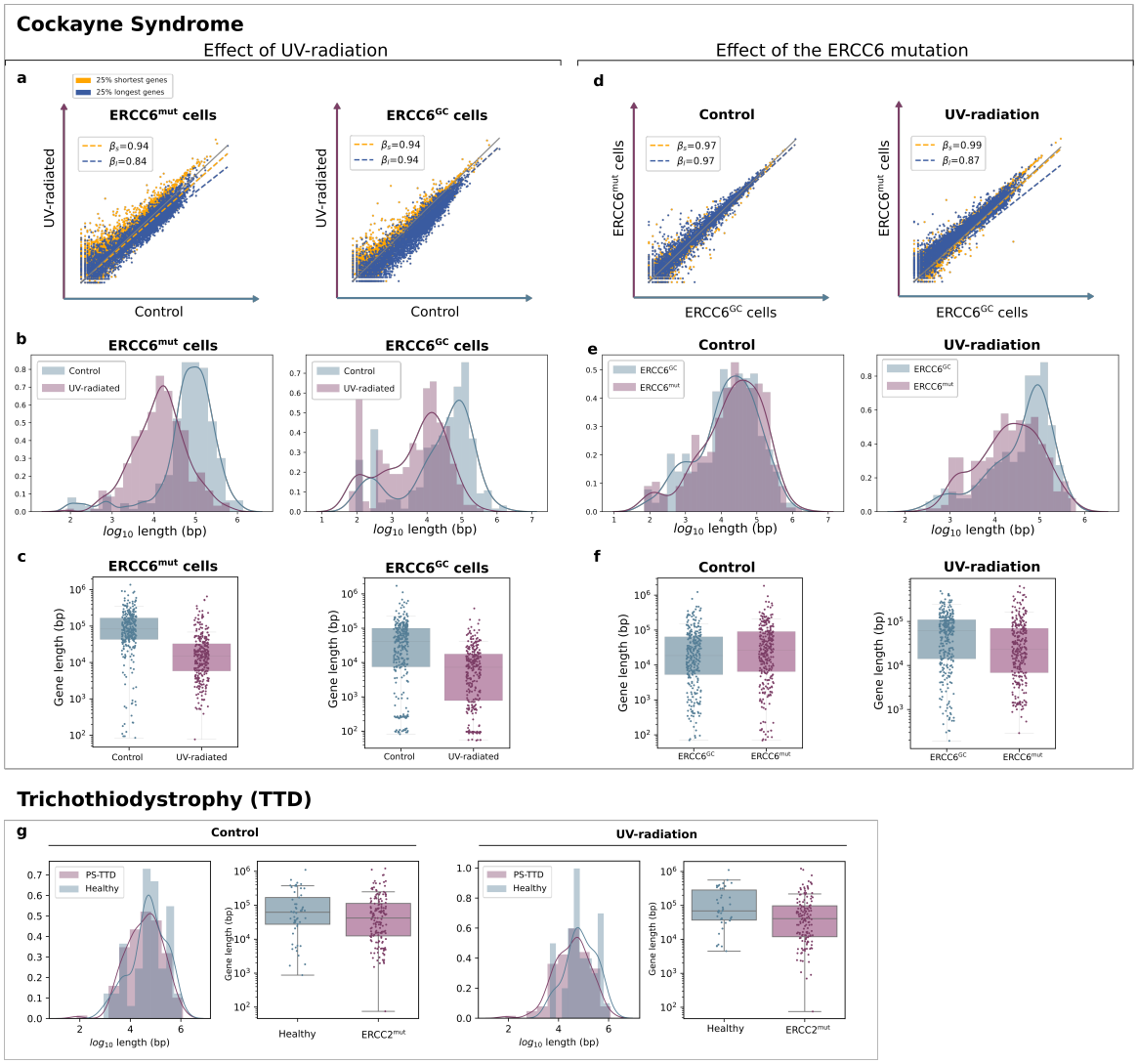


Figure 7. Human Cockayne Syndrome-and trichothiodystrophy (TTD)-derived cells mimic age-associated decrease in the expression of long genes. a-c, Effect of UV-radiation on cells carrying a mutation in Cockayne Syndrome group B (ERCC6). Average gene expression in UV-radiated cells *vs* in normal conditions in mutant ($ERCC^{mut}$) and gene corrected ($ERCC^{GC}$) cells (**a**). The gene lengths of the 300 most differentially expressed genes between cells with and without UV-radiation exposure in mutant and gene corrected cells, shown as overlapped density plots (**b**) and separate boxplots (**c**). **d-f**, Baseline effect of the ERCC6 mutation on length-dependent expression. Average gene expression between mutant ($ERCC^{mut}$) and gene corrected ($ERCC^{GC}$) cells in normal conditions (control) and after UV-radiation exposure (**d**). Length of the 300 most differentially expressed genes between mutant and gene corrected cells, shown as overlapped density plots (**e**) and separate boxplots (**f**). **g**, Length of the DEGs ($|\ logFC\ | \geq 2$ and p-value ≤ 0.05) between a PS-TTD patient and her healthy mother in basal conditions (control) and upon UV-radiation.

been applied to these genes. We analyzed the length of GAGs (Figure 8) and found that genes that are downregulated with aging tend to be longer than those that were found to be upregulated, and that their difference in length is statistically significant (Figure 8a, Wilcoxon-Mann-Whitney test, $p - value < 0.01$).

In humans, the first large-scale meta-analysis (14,983 individuals) of aging-related gene expression profiles identified 1,497 genes differentially expressed with chronological age in peripheral blood mononuclear cells [38]. Interestingly, long genes downregulated with aging in this human cohort,

the differences in length between upregulated and downregulated genes being statistically significant (Figure 8b). Overall, these data suggest that transcriptomic aging signatures are influenced by gene length-dependent transcriptional decay.

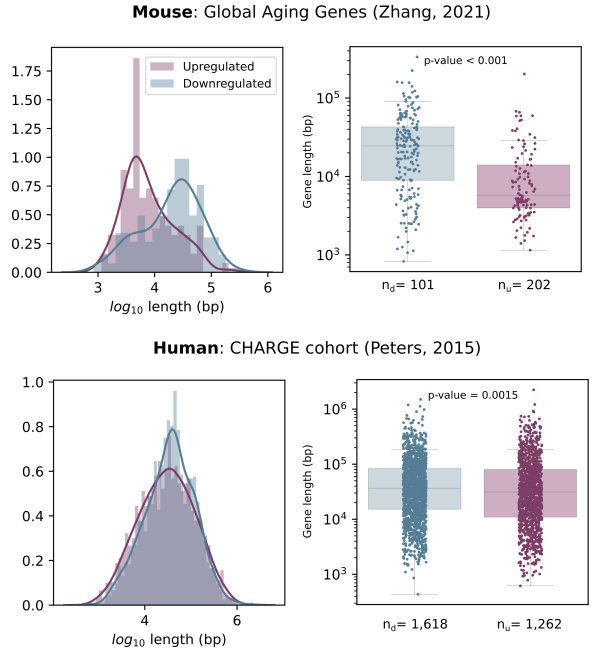


Figure 8. Down-regulated genes are longer than up-regulated genes in two published aging transcriptomic signatures. The length of the genes from two aging signatures (murine and human) are shown as two overlapped histograms and separate boxplots. The number of up-and down-regulated genes in each signature are shown as n_d and n_u, respectively. The gene length is different between the two categories according to the Mann-Whitney test (p-values shown in the figure).

Discussion

In this article, we report that a generalized age-related decline in gene expression is dependent on gene length. The fact that gene length affects mRNA expression levels has long been known [39]. In early development, gene size and architecture influences the expression timing of specific genes [40]. This is also true more generally, for instance in the immediate cellular response to external stimuli, where shorter pre-mRNA molecules are synthesized first [41]. Furthermore, gene lengths appear to be compartmentalized among chromosomes, and tissue-specific expression patterns may be detected [42].

RNA polymerase II (RNAP II)-driven transcription can be divided into initiation, pausing, elongation, 3' end formation and termination stages; each step being tightly regulated [43]. Once initiated, transcription pauses downstream from the transcription start site and requires specific signaling for pause-release, elongation and processivity. Cyclin-dependent kinases CDK12 and CDK13 seem to be involved in the regulation of RNAP II elongation, processivity and selection of alternative polyadenylation sites [44]. Of interest, the GC content of the initially transcribed sequence determines early RNAP II elongation rates, and recognition of a 5' splice site (SS) by U1 snRNP promotes RNAP II elongation potential [45]. This is related to a process known as *telescripting*, whereby U1 snRNP base pairing with 5'SS avoids premature 3' end cleavage and polyadenylation at cryptic intronic sites [46, 47]. It is likely that long gene transcription is mediated by many other RNA-binding proteins (RBPs) as well, many of which have additional functions in the regulation of pre-mRNA splicing [48]. In fact, only about

half of the introns present in newly synthesized pre-mRNA are co-transcriptionally spliced [49], further supporting alternative roles for specific RBP subsets. Although we have no mechanistic understanding of which dysfunction is mediating the apparent loss of long gene transcription associated to aging, our data may generate new avenues for aging-related research, where the relevance of pathways related to RNAP II elongation and processivity remains virtually unexplored.

Premature transcript termination by RNAP II has already been described in some contexts. An increase in elongation rate (speed) concomitant to premature termination at cryptic intronic polyadenylation signals has recently been reported during heat shock, which was mediated by inhibition of U1 telescripting [50]. Interestingly, failure to target the stalled RNAP II for degradation by polyubiquitination of a single residue is enough to shutdown long gene transcription, the expression of shorter genes being unaffected [51, 52]. Further, the concept of *long-gene transcriptopathy* has been proposed as a possible mechanism underlying a number of neurological and psychiatric disorders, some of which are age-associated [53, 54, 48]. RNA-binding protein SFPQ mediates CDK9 recruitment to the transcription elongation complex, which activates RNAP II-CTD. Neuron-specific ablation of SFPQ downregulated a regulon of 135 genes, which account for less than 10 percent of the genes with a pre-mRNA >100 kb in length, inducing neuronal cell death and embryonic lethality [54]. Similarly, muscle-specific ablation of SFPQ induced metabolic myopathy, severe progressive muscle mass reduction and impairment of motor function. This was shown to be mediated by downregulation of long genes regulating energy metabolism in skeletal muscle [48]. While the specific mechanisms underlying the generalized age-associated downregulation of long genes that we report here remain to be determined, it seems likely that they will be related to some of the aforementioned mechanisms. For example, a longitudinal analysis of gene expression differences in a human cohort that followed 65 healthy individuals between ages 70 and 80 [55] found changes in the expression of the *SFPQ* gene among the strongest associations with age. Of note, the key importance of RNA metabolism dysregulation in human aging has long been known [56].

Accumulation of genotoxic damage with chronological age is pervasive, and it may also be significantly incremented through lifestyle choices [27, 31, 57, 58]. The fact that augmented DNA damage specifically induces downregulation of long genes is of great interest. A recent study has shown that UV-mediated global transcription shutdown favored transcription restart from shorter mRNAs with less exons [59]. Similarly, transcription blockage by DNA damage is known to generate neurodegenerative processes associated to human genetic syndromes deficient in nucleotide excision repair, such as Cockayne Syndrome and xeroderma pigmentosum [60]. Our data showing that several models of progeroid disease specifically downregulate long genes are most likely true as well for other TC-NER syndromes.

The search for aging-related gene signatures has provided relatively little advance to the field. In our opinion, the straightforward mechanism depicted here (of DNA damage-induced loss of RNAP II processivity as a molecular driver of aging) might better explain many of the age-associated features and may thus provide a fruitful research avenue for the aging field. Future work should shed light on the specific mechanisms underlying loss of long gene transcription associated with aging.

Methods

Data inclusion criteria

In order to analyze balanced aging datasets, samples were selected according to the following criteria: 1) When sex annotations where available, same-sex datasets were generated. 2) Individuals of the

same age were used to create the "young" and the "old" cohorts. 3) In datasets including samples from different sub-tissues, samples corresponding to the sub-tissues with representation in the two age cohorts were selected.

In murine datasets derived from *Tabula Muris Senis* data, 3 month-old and 24 month old mice were used to form the young and old cohorts, respectively. In all *TMS* female murine aging datasets 18 month animals were used to form the old cohort. In the murine dermal fibroblast dataset [21], samples from newborn mice were not included.

Regarding human aging datasets, samples from newborn and middle-aged individuals were discarded and sex-stratified cohorts where created when possible. In the human aging pancreas dataset [22], samples from pediatric donors as well as those from a 38-year old patient were removed. Thus, only two young (21 and 22 years old) and two old (44 and 54 years old) donors were included in the aging dataset.

In the human trachea of heavy smokers and never-smokers dataset [32] only donors aged over 50 years were included in the dataset to avoid age as a confounding variable.

Data processing pipeline

Single-cell RNA-seq datasets were preprocessed using a standard preprocessing pipeline in *Scanpy* [61]: normalization, log-transformation of counts, feature selection using *triku* [62], dimensionality reduction through Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection (UMAP) [63], and community detection using *Leiden* [64]. In some cases, when the original labels were too granular, some cell identities were merged into broader categories before proceeding to downstream analyses.

Datasets

Male murine aging datasets

TMS male mice aged 3 months and 24 months were selected to create balanced datasets of aging of 11 organs (12 comparisons): bladder, brain, brain myeloid, heart, kidney, liver, lung, muscle, pancreas, skin, spleen and thymus (Almanzar et al. [17]).

Female murine aging datasets

Due to the lack of available 24 month-old females in the *TMS* dataset, we chose a set of 3 month and 18 month-old mice to create 12 balanced female aging datasets: TMSF muscle, TMSF brain, TMSF brain myeloid, TMSD heart, TMSF heart, TMSF thymus, TMSF skin, TMSF pancreas, TMSD mammary gland, TMSF mammary gland, TMSF spleen and TMSF kidney.

Additional murine and human datasets

We analyzed six additional murine aging datasets of several tissues: lung cells from 3 and 24 month old mice (Angelidis et al. [18], GEO accession GSE124872), lung, spleen and kidney cells from 7 and 21 months old mice (Kimmel et al. [19], GSE132901), brain cells from 2-3 and 21-23 month old mice (Ximerakis et al. [20], GSE129788) and dermal fibroblasts from 2 and 18 month old mice (Salzer et al. [21], GSE111136). We also analyzed four human datasets: lung cells from 46 and 75 years old male healthy donors (Travaglini et al. [24], available at Synapse under accession syn21041850), lung cells from young (21, 22, 32, 35 and 41 years old) and old (64, 65, 76 and 88 years old) male and female

healthy donors Raredon et al. [25], GSE133747), pancretic cells from X and Y years old male and female healthy donors (Enge et al. [22], GSE81547), and whole-skin cells from X and Y years old donors ([23], GSE130973). Murine lung, human lung and human pancreas datasets were processed and cell type annotated as in Ibáñez-Solé et al. [65].

Murine aging heart

Four aging balanced datasets were created from samples from the *TMS* FACS heart and the *TMS* droplet heart and aorta datasets. All mice aged 3 months, 18 months and 21 months were selected and combined so that all mice representing an age cohort within a dataset were of equal age and sex: TMS FACS male (3-24 months), TMS FACS female (3-18 months), TMS droplet female (3-18 months) and TMS droplet female (3-21 months).

Murine UV-radiated skin

The datasets corresponding to the three conditions (*healthy*, *UV-radiated* and *vitamin D*) were downloaded from the Gene Expression Omnibus (GSE173385). We checked that the age of the mice used in the study was identical between conditions. The three datasets were subjected to the standard processing pipeline described in Data processing pipeline separately. Then, the Leiden community detection algorithm was run and cell type annotations were added to the resulting clusters based on the expression of known cell type markers. The murine dermal cell type characterization by Joost et al. was used as a reference.

The clusters were annotated based on the following gene markers: «IFE basal» (basal keratinocytes from the interfollicular epidermis, *Krt5, Krt14, Mt2*); «IFE diff.» (differentiating keratinocytes, *Krt1, Krt10, Ptgs1*); «IFE kerat.» (terminally differentiated cells in the keratinyzed layer, *Lor, Flg2.*); «HF» (hair follicle cells, *Krt17, Krt79, Sox9*); «Fibroblast» (*Col1a1, Col3a1, Col1a2, Dcn, Lum, Sparc*); «Myeloid» (*Cd74, Lyz2*); «SG» (sebaceous gland cells, *Mgst1, Scd1, Krt25, Pparg*); «T cell» (*Cd3d, Thy1, Nkg7*); «EC» (endothelial cells, *Mgp, Fabp4*); «Melanocyte» (*Mlana, Pmel, Tyrp1*); «Erythrocyte» (*Hbb-bs, Hbb-bt, Hbba-a2*).

The Lilliefors normality test [67] was conducted on the log-transformed lengths of the differentially expressed genes for each of the conditions, using Python module `statsmodel`. The null hypothesis – that the log_{10} gene lengths follow a normal distribution – could not be rejected (cutoff: 0.05), meaning that the distribution of gene lengths within each group is normally distributed. We tested whether the mean lengths of the DEGs were significantly different across conditions using ANOVA (`stats.f_oneway`). The null hypothesis that the three means were equal was rejected (p-value 3.67E-06). Post-hoc analysis (Tukey test, `scikit_posthocs.posthoc_tukey`) was run to test which of the pairwise comparisons between the three conditions yielded a statistically significant difference. Additionally, statistical significance was confirmed with non-parametric alternatives: Kruskal-Wallis (`stats.kruskal`) and Dunn test (`scikit_posthocs.posthoc_dunn`).

Human airway cells from heavy smokers

The dataset used in Goldfarbmuren et al. was downloaded from the Gene Expression Omnibus (GSE134174). Original cell type annotations were used, but subtypes of the same cell types were pooled into a single category. The final dataset contained 13 cell types: «Diff. basal» (differentiating basal cells), «Prolif. basal» (proliferating basal cells), «Prot. basal » (proteasomal basal cells), «ciliated» (the two mature ciliated clusters –A and B– were pooled together), «ionocytes», «PNEC» (pulmonary neuroendocrine cells), «secretory/ciliating» (hybrid secretory early ciliating cells), «KRT8 high»,

«secretory» (mucus secretory cells), «tuft-like» (Tuft-like cells), «SMG basal» (basal cells from the submucosal gland or SMG, the two clusters –A and B– were pooled into a single category), «SMG myoepithelial» (myoepithelial cells from the SMG), «SMG secretory» (mucus secretory cells from the SMG).

In order to control for age as a possible confounding factor, we checked the ages of the subjects in the original dataset. We discarded the youngest donors and only kept samples from donors aged >50 years. The final dataset consisted of 21,425 cells from 8 donors. Heavy smokers (*T101, T120, T154, T167, T85*) were aged 55-66 years, and never-smokers (*T164, T165, T166*) were 64-68 years old. Since the average never-smoker age is slightly higher than the average heavy-smoker age, we can safely attribute transcriptional changes between these two groups to their smoking status.

The Lilliefors test was used to test whether the log_{10} length of the DEGs for the two conditions ("heavy smokers" and "never-smokers") were normally distributed. The null hypothesis could be rejected (cut-off: 0.05) for the "never-smokers", meaning that DEGs associated with that condition were not normally distributed, so a MannWhitney U test was used to compare between the means of the two distributions.

Effect of ERCC6 mutation of susceptibility to UV-radiation

The dataset by Wang et al. was downloaded from the Gene Expression Omnibus (GSE124208). The following samples were included in the dataset: *GSM3525718*, *GSM3525717*, *GSM3525714*, *GSM3525715*, *GSM3525719*, *GSM3525716*, *GSM3525713* and *GSM3525720*. Those samples correspond to four experimental conditions: MSCs from Cockayne syndrome patients carrying the ERCC6 mutation, with (*UV*) and without (*ct*) UV-radiation treatment (*MSC_mut_ct*, *MSC_mut_UV*); MSCs from gene-corrected cells with and without UV radiation treatment (*MSC_GC_ct* and *MSC_GC_UV*). All samples were merged into a single dataset and expression values were log-transformed.

Effect of ERCC2 mutation of susceptibility to UV-radiation

The complete list of DEGs between a cancer-free PS-TTD patient carrying a mutated ERCC2 gene and her healthy mother in basal conditions and upon UV-radiation were obtained from the Supplementary Material provided by Lombardi et al. [37]. From the original DEG list, we selected the genes with a log fold-change greater than 2 (either overexpressed in the sample from the PS-TTD patient or in the sample from the healthy donor). The same threshold for statistical significance (p-value $\leq$0.05) as the one used by the original authors was used.

Gene length analysis

Human and mouse gene length annotations for were obtained from Biomart. Total gene length was calculated as the difference between the transcription end site and the transcription start site.

Length-dependent difference in expression in aging and genotoxic conditions

Two different types of analysis were run between conditions: global average gene expression and length-dependence of transcriptional decay and gene length analysis of the differentially expressed genes between conditions.

Gene length dependence in age-related transcriptional decay

Here, we computed the average gene expression across all cells for a pair of conditions (for instance, "young" and "old"). We used a scatter plot to represent each gene according to its average expression in old cells (y axis) against its average expression in young cells (x axis). This is a way of looking at how predictable the expression of each particular gene is in old cells based on the expression of the same gene in young cells. As we observed that most genes show a great correlation between young and old cells, even though many of them show expression levels that are lower than what we would have expected from their expression in young individuals, we then looked at the role gene length plays in this transcriptional decay. We did so by splitting the transcriptome into four quartiles according to their length. we considered whole sequence length from transcription start site to transcription end site. Then, we fitted a linear regression model to the average gene expression in old and young cells for each of the quartiles, thus obtaining a separate linear model for each quartile, using the formula $ME_{old} \sim ME_{young} * Q$, where ($ME_{old}$ and ME_{young} are the mean expression vectors for old and young cells, and Q is the vector that assigns each gene to a length quartile, to be used as a factor by the linear model). We observed that the shorter the genes included in the linear model (for instance, Q1 genes), the greater was the slope of the resulting straight. We performed statistical analysis to compare between the slope of the Q1 model against each of the three remaining models (Q2, Q3 and Q4).

The same analysis was extended to conditions other than aging, by making analogous comparisons. In the UV-radiated murine skin analysis, we compared UV-radiated skin against the healthy skin control (to test for the effect of UV-radiation), the UV-radiated skin against the vitamin D-treated and UV-radiated skin (effect of vitamin D treatment on damage caused by UV-radiation), and the vitamin D-treated skin against the healthy skin control (effect of UV-radiation after vitamin D treatment). In the analysis on the murine model for Cockayne syndrome we compared between each of the knock outs ($Adh5^{-/-}$, $Csb^{m/m}$, and double KO) against the wild type (WT). In the analysis of human mesenchymal stromal cells derived from Cockayne syndrome patients, we compared between the following conditions: UV-radiated cells against control (both in mutant and gene corrected cells), and $ERCC^{mut}$ against $ERCC^{GC}$ (to test for the effect of carrying the ERCC6 mutation, both in normal conditions and after UV-radiation exposure).

Gene length analysis of the differentially expressed genes between conditions

We carried out two types of differential expression analysis: overall differential expression between conditions and differential expression at the cell type level.

Overall differential expression between conditions is based on the assumption that the changes in cell type composition between the conditions to be compared are negligible, so that the genes that are detected to be differentially expressed do not correspond to markers defining specific cell types that are more abundant in one of the conditions. Differential expression analysis between conditions at the cell type level identifies genes that are over-expressed in one of the conditions. Of course, DEGs can only be computed for cell types that are present in the conditions to be compared in sufficient amounts (we used 10 cells as the minimum). Its output is not directly affected by changes in cell type composition between conditions. However, if the abundance of cell type under study is very different between conditions – if one cell type is very rare in one of the conditions – the population might not be well sampled for that condition and the gene length analysis might not be reliable. We therefore use the two approaches as they are complementary to one another. In either case, we used the Scanpy function `sc.tl.rank_genes_groups` with `method = "wilcoxon"` to obtain the top 300 differentially

expressed genes between conditions.

In most cases, pairwise comparisons were made, as in the aging analysis ("young" *vs* "old") or when analyzing the effect of smoking of human airways ("never-smokers" *vs* "heavy smokers"). In those cases, two lists of genes were obtained: one per condition. In the analysis of murine UV-radiated skin (Figure 4), we compared between the three conditions simultaneously. In that case, each of three DEG lists corresponds to the genes that are over-expressed in one condition against the other two conditions pooled together.

First, the Lilliefors test was used check whether gene lengths in each of the conditions were normally distributed. In cases where the null hypothesis could be rejected (p-value < 0.05) in at least one of the conditions to be compared, a non parametric test was used to compare between means. In order to make statistical comparisons between the mean gene length between conditions, we used the following tests: Student's T test (two conditions, normally distributed), Mann-Whitney's U test (two conditions, not normally distributed), ANOVA (three conditions) and Tukey's test for post-hoc analysis.

Code availability

Jupyter notebooks and R scripts for reproducing the analyses can be found in GitLab.

References

1. Björn Schumacher, Joris Pothof, Jan Vijg, and Jan H. J. Hoeijmakers. The central role of dna damage in the ageing process. *Nature*, 592:695–703, 2021. doi: 10.1038/s41586-021-03307-7.

2. Matt Yousefzadeh, Chathurika Henpita, Rajesh Vyas, Carolina Soto-Palma, Paul Robbins, and Laura Niedernhofer. Dna damage - how and why we age? *eLife*, 10:e62852, 2021. doi: 10.7554/eLife.62852.

3. Stephen Frenk and Jonathan Houseley. Gene expression hallmarks of cellular ageing. *Biogerontology*, 19(6):547566, 2018. doi: 10.1007/s10522-018-9750-z.

4. João Pedro de Magalhães, João Curado, and George M. Church. Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics*, 25(7):875881, 2009. doi: 10.1093/bioinformatics/btp073.

5. Robi Tacutu, Daniel Thornton, Emily Johnson, Arie Budovsky, Diogo Barardo, Thomas Craig, Eugene Diana, Gilad Lehmann, Dmitri Toren, Jingwei Wang, and et al. Human ageing genomic resources: New and updated databases. *Nucleic Acids Research*, 46(D1), 2017. doi: 10.1093/nar/gkx1042.

6. Daniel Palmer, Fabio Fabris, Aoife Doherty, Alex A. Freitas, and João Pedro de Magalhães. Ageing transcriptome meta-analysis reveals similarities between key mammalian tissues. *Aging*, 2019. doi: 10.1101/815381.

7. W. Zhang, S. Zhang, P. Yan, J. Ren, M. Song, J. Li, J. Lei, H. Pan, S. Wang, X. Ma, and et al. A single-cell transcriptomic landscape of primate arterial aging. *Nature Communications*, 11(1), 2020. doi: 10.1038/s41467-020-15997-0.

8. R. Stegeman and V.M. Weake. Transcriptional signatures of aging. *Journal of Molecular Biology*, 429(16):24272437, 2017. doi: 10.1016/j.jmb.2017.06.019.

9. Daniel J. Haustead, Andrew Stevenson, Vishal Saxena, Fiona Marriage, Martin Firth, Robyn Silla, Lisa Martin, Katharine F. Adcroft, Suzanne Rea, Philip J. Day, and et al. Transcriptome analysis of human ageing in male skin shows mid-life period of variability and central role of nf-b. *Scientific Reports*, 6(1), 2016. doi: 10.1038/srep26846.

10. Anthony Yannarell, Dorothy E. Schumm, and Thomas E. Webb. Age-dependence of nuclear rna processing. *Mechanisms of Ageing and Development*, 6:259264, 1977. doi: 10.1016/0047-6374(77)90026-4.

11. Hannes Lans, Jan H. Hoeijmakers, Wim Vermeulen, and Jurgen A. Marteijn. The dna damage response to transcription stress. *Nature Reviews Molecular Cell Biology*, 20(12):766784, 2019. doi: 10.1038/s41580-019-0169-4.

12. Lea H. Gregersen and Jesper Q. Svejstrup. The cellular response to transcription-blocking dna damage. *Trends Biochem Sci.*, 43(5):327–341, 2018. doi: 10.1016/j.tibs.2018.02.010.

13. Jianming Wang, Martina Muste Sadurni, and Marco Saponaro. Rnapii response to transcriptionblocking dna lesions in mammalian cells. *The FEBS Journal*, 2022. doi: 10.1111/febs.16561.

14. Sourena SoheiliNezhad, Robert J. Linden, Marcel Olde Rikkert, Emma Sprooten, and Geert Poelmans. Long genes are more frequently affected by somatic mutations and show reduced expression in alzheimer's disease: Implications for disease etiology. *Alzheimer's & Dementia*, 17(3):489499, 2020. doi: 10.1002/alz.12211.

15. Inês Lopes, Gulam Altab, Priyanka Raina, and João Pedro de Magalhães. Gene size matters: An analysis of gene length in the human genome. *Frontiers in Genetics*, 12, 2021. doi: 10.3389/fgene.2021.559998.

16. W. P. Vermeij, M. E. T. Dollé, E. Reiling, D. Jaarsma, C. Payan-Gomez, C. R. Bombardieri, H. Wu, A. J. M. Roks, S. M. Botter, B. C. van der Eerden, S. A. Youssef, R. V. Kuiper, B. Nagarajah, C. T. van Oostrom, R. M. C. Brandt, S. Barnhoorn, S. Imholz, J. L. A. Pennings, A. de Bruin, A. Gyenis, J. Pothof, J. Vijg, H. van Steeg, and J. H. J. Hoeijmakers. Restricted diet delays accelerated ageing and genomic stress in dna-repair-deficient mice. *Nature*, 537:427–431, 2016. doi: 10.1038/nature19329.

17. N. Almanzar, J. Antony, and the Tabula Muris Consortium. A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature*, 583:590–595, 2020. doi: 10.1038/s41586-020-2496-1.

18. I. Angelidis, L. M. Simon, and I. E. et al. Fernandez. An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *Nature Communications*, 10, 2019. doi: https://doi.org/10.1038/s41467-019-08831-9.

19. Jacob C. Kimmel, Lolita Penland, Nimrod D. Rubinstein, David G. Hendrickson, David R. Kelley, and Adam Z. Rosenthal. Murine single-cell rna-seq reveals cell-identity- and tissue-specific trajectories of aging. *Genome Research*, 29(12):20882103, 2019. doi: 10.1101/gr.253880.119.

20. M. Ximerakis, S. L. Lipnick, B. T. Innes, S. K. Simmons, X. Adiconis, D. Dionne, B. A. Mayweather, L. Nguyen, Z. Niziolek, C. Ozek, V. L. Butty, R. Isserlin, S. M. Buchanan, S. S. Levine, A. Regev, G. D. Bader, J. Z. Levin, and L. L. Rubin. Single-cell transcriptomic profiling of the aging mouse brain. *Nat Neurosci*, 22(10), 2019. doi: https://doi:10.1038/s41593-019-0491-3.

21. M. C. Salzer, A. Lafzi, A. Berenguer-Llergo, C. Youssif, A. Castellanos, G. Solanas, F. Oliveira Peixoto, C. S-O. Attolini, N. Prats, M. Aguilera, and et al. Identity noise and adipogenic traits characterize dermal fibroblast aging. *Cell*, 175(6), 2018. doi: 10.1016/j.cell.2018.10.012.

22. M. Enge, H. E. Arda, M. Mignardi, J. Beausang, R. Bottino, S. K. Kim, and S. R. Quake. Single-cell analysis of human pancreas reveals transcriptional signatures of aging and somatic mutation patterns. *Cell*, 171(2), 2017. doi: https://doi:10.1016/j.cell.2017.09.004.

23. L. Solé-Boldo, G. Raddatz, and S. et al. Schütz. Single-cell transcriptomes of the human skin reveal age-related loss of fibroblast priming. *Commun Biol*, 3(188), 2020. doi: https://doi.org/10.1038/s42003-020-0922-4.

24. Kyle J. Travaglini, Ahmad N. Nabhan, Lolita Penland, Rahul Sinha, Astrid Gillich, Rene V. Sit, Stephen Chang, Stephanie D. Conley, Yasuo Mori, Jun Seita, Gerald J. Berry, Joseph B. Shrager, Ross J. Metzger, Christin S. Kuo, Norma Neff, Irving L. Weissman, Stephen R. Quake, and Mark A. Krasnow. A molecular cell atlas of the human lung from single-cell rna sequencing. *Nature*, 587:619625, 2020. doi: 10.1038/s41586-020-2922-4.

25. Micha Sam Brickman Raredon, Taylor Sterling Adams, Yasir Suhail, Jonas Christian Schupp, Sergio Poli, Nir Neumark, Katherine L. Leiby, Allison Marie Greaney, Yifan Yuan, Corey Horien, and et al. Single-cell connectomic analysis of adult mammalian lungs. *Science Advances*, 5(12), 2019. doi: 10.1126/sciadv.aaw3851.

26. Martin Jinye Zhang, Angela Oliveira Pisco, Spyros Darmanis, and James Zou. Mouse aging cell atlas analysis reveals global and cell type-specific aging signatures. *eLife*, 10, 2021. doi: 10.7554/elife.62293.

27. Iñigo Martincorena, Amit Roshan, Moritz Gerstung, Peter Ellis, Peter Van Loo, Stuart McLaren, David C Wedge, Anthony Fullam, Ludmil B Alexandrov, Jose M Tubio, Lucy Stebbings, Andrew Menzies, Sara Widaa, Michael R Stratton, Philip H Jones, and Peter J Campbell. Tumor evolution. high burden and pervasive positive selection of somatic mutations in normal human skin. *Science*, 348:880–886, 2015. doi: 10.1126/science.aaa6806.

28. T Passeron, HW Lim, CL Goh, HY Kang, F Ly, A Morita, J Ocampo Candiani, S Puig, S Schalka, L Wei, B Dréno, and J. Krutmann. Photoprotection according to skin phototype and dermatoses: practical recommendations from an expert panel. *J. Eur. Acad. Dermatol. Venereol.*, 35:1460–1469, 2021. doi: 10.1111/jdv.17242.

29. Clare Gordon-Thomson, Wannit Tongkao-on, Eric J. Song, Sally Carter, Katie M. Dixon, and Rebecca S. Mason. Protection from ultraviolet damage and photocarcinogenesis by vitamin d compounds. *Adv Exp Med Biol.*, 810:303–328, 2014. doi: 10.1007/978-1-4939-0437-2_17.

30. Yuanbin Lin, Zhanglei Cao, Tianqi Lyu, Tong Kong, Qian Zhang, Kerong Wu, Yuhui Wang, and Jianping Zheng. Single-cell rna-seq of uvb-radiated skin reveals landscape of photoaging-related inflammation and protection by vitamin d. *Gene*, 831:146563, 2022. doi: 10.1016/j.gene.2022.146563.

31. Zhenqiu Huang, Shixiang Sun, Moonsook Lee, Alexander Y Maslov, Miao Shi, Spencer Waldman, Ava Marsh, Taha Siddiqui, Xiao Dong, Yakov Peter, Ali Sadoughi, Chirag Shah, Kenny Ye, Simon D Spivack, and Jan Vijg. Single-cell analysis of somatic mutations in human bronchial

epithelial cells in relation to aging and smoking. *Nature Genetics*, 54:492–498, 2022. doi: 10.1038/s41588-022-01035-w.

32. Katherine C. Goldfarbmuren, Nathan D. Jackson, Satria P. Sajuthi, Nathan Dyjack, Katie S. Li, Cydney L. Rios, Elizabeth G. Plender, Michael T. Montgomery, Jamie L. Everman, Preston E. Bratcher, and et al. Dissecting the cellular specificity of smoking effects and reconstructing lineages in the human airway epithelium. *Nature Communications*, 11(1), 2020. doi: 10.1038/s41467-020-16239-z.

33. Matthias Rieckher, George A Garinis, and Björn Schumacher. Molecular pathology of rare progeroid diseases. *Trends Mol Med.*, 27:907–922, 2021. doi: 10.1016/j.molmed.2021.06.011.

34. Carla Umansky, Agustín E Morellato, Matthias Rieckher, Marco A Scheidegger, Manuela R Martinefski, Gabriela A Fernández, Oleg Pak, Ksenia Kolesnikova, Hernán Reingruber, Mariela Bollini, Gerry P Crossan, Natascha Sommer, María Eugenia Monge, Björn Schumacher, and Lucas B Pontel. Endogenous formaldehyde scavenges cellular glutathione resulting in redox disruption and cytotoxicity. *Nat Commun.*, 13:745, 2022. doi: 10.1038/s41467-022-28242-7.

35. Lee Mulderrig, Juan I Garaycoechea, Zewen K Tuong, Christopher L Millington, Felix A Dingler, John R Ferdinand, Liam Gaul, John A Tadross, Mark J Arends, Stephen O'Rahilly, Gerry P Crossan, Menna R Clatworthy, and Ketan J Patel. Aldehyde-driven transcriptional stress triggers an anorexic dna damage response. *Nature*, 600:158–163, 2021. doi: 10.1038/s41586-021-04133-7.

36. Si Wang, Zheying Min, Qianzhao Ji, Lingling Geng, Yao Su, Zunpeng Liu, Huifang Hu, Lixia Wang, Weiqi Zhang, Keiichiro Suzuiki, Yu Huang, Puyao Zhang, Tie-Shan Tang, Jing Qu, Yang Yu, Guang-Hui Liu, and Jie Qiao. Rescue of premature aging defects in cockayne syndrome stem cells by crispr/cas9-mediated gene correction. *Protein Cell*, 11:1–22, 2020. doi: 10.1007/s13238-019-0623-2.

37. Anita Lombardi, Lavinia Arseni, Roberta Carriero, Emmanuel Compe, Elena Botta, Debora Ferri, Martina Uggè, Giuseppe Biamonti, Fiorenzo A. Peverali, Silvia Bione, and et al. Reduced levels of prostaglandin i2 synthase specify the cancer-free trichothiodystrophy. *Proceedings of the National Academy of Sciences*, 118(26), 2021. doi: 10.1073/pnas.2024502118.

38. Marjolein J. Peters, Roby Joehanes, Luke C. Pilling, Claudia Schurmann, Karen N. Conneely, Joseph Powell, Eva Reinmaa, George L. Sutphin, Alexandra Zhernakova, Katharina Schramm, Yana A. Wilson, Sayuko Kobes, Taru Tukiainen, NABEC/UKBEC Consortium, Yolande F. Ramos, Harald H. H. Göring, Myriam Fornage, Yongmei Liu, Sina A. Gharib, Barbara E. Stranger, Philip L. De Jager, Abraham Aviv, Daniel Levy, Joanne M. Murabito, and Peter J. et al. Munson. The transcriptional landscape of age in human peripheral blood. *Nat Commun*, 6:8570, 2015. doi: 10.1038/ncomms9570.

39. Francesca Chiaromonte, Webb Miller, and Eric E Bouhassira. Gene length and proximity to neighbors affect genome-wide expression levels. *Genome Res.*, 13:2602–2608, 2003. doi: 10.1101/gr.1169203.

40. Patricia Heyn, Alex T Kalinka, Pavel Tomancak, and Karla M Neugebauer. Introns and gene expression: cellular constraints, transcriptional regulation, and evolutionary consequences. *Bioessays*, 37:148–154, 2015. doi: 10.1002/bies.201400138.

41. Killeen S Kirkconnell, Brian Magnuson, Michelle T Paulsen, Brian Lu, Karan Bedi, and Mats Ljungman. Gene length as a biological timer to establish temporal transcriptional regulation. *Cell Cycle*, 16:259–270, 2017. doi: 10.1080/15384101.2016.1234550.

42. Jay C Brown. Role of gene length in control of human gene expression: chromosome-specific and tissue-specific effects. *Int. J. Genomics*, 2021:8902428, 2021. doi: 10.1155/2021/8902428.

43. Patrick Cramer. Organization and regulation of gene transcription. *Nature*, 573:45–54, 2019. doi: 10.1038/s41586-019-1517-4.

44. Zheng Fan, Jennifer R Devlin, Simon J Hogg, Maria A Doyle, Paul F Harrison, Izabela Todorovski, Leonie A Cluse, Deborah A Knight, Jarrod J Sandow, Gareth Gregory, Andrew Fox, Traude H Beilharz, Nicholas Kwiatkowski, Nichollas E Scott, Ana Tufegdzic Vidakovic, Gavin P Kelly, Jesper Q Svejstrup, Matthias Geyer, Nathanael S Gray, Stephin J Vervoort, and Ricky W Johnstone. Cdk13 cooperates with cdk12 to control global rna polymerase ii processivity. *Sci Adv.*, 6:eaaz5041, 2020. doi: 10.1126/sciadv.aaz5041.

45. Hanneke Vlaming, Claudia A Mimoso, Andrew R Field, Benjamin J E Martin, and Karen Adelman. Screening thousands of transcribed coding and non-coding regions reveals sequence determinants of rna polymerase ii elongation potential. *Nat Struct Mol Biol.*, 29:613–620, 2022. doi: 10.1038/s41594-022-00785-9.

46. Daisuke Kaida, Michael G Berg, Ihab Younis, Mumtaz Kasim, Larry N Singh, Lili Wan, and Gideon Dreyfuss. U1 snrnp protects pre-mrnas from premature cleavage and polyadenylation. *Nature*, 468:664–668, 2010. doi: 10.1038/nature09479.

47. Michael G Berg, Larry N Singh, Ihab Younis, Qiang Liu, Anna Maria Pinto, Daisuke Kaida, Zhenxi Zhang, Sungchan Cho, Scott Sherrill-Mix, Lili Wan, and Gideon Dreyfuss. U1 snrnp determines mrna length and regulates isoform expression. *Cell*, 150:53–64, 2012. doi: 10.1016/j.cell.2012.05.029.

48. Motoyasu Hosokawa, Akihide Takeuchi, Jun Tanihata, Kei Iida, Shin'ichi Takeda, and Masatoshi Hagiwara. Loss of rna-binding protein sfpq causes long-gene transcriptopathy in skeletal muscle and severe muscle mass reduction with metabolic myopathy. *iScience*, 13:229–242, 2019. doi: 10.1016/j.isci.2019.02.023.

49. Karan Bedi, Brian R Magnuson, Ishwarya Narayanan, Michelle Paulsen, Thomas E Wilson, and Mats Ljungman. Co-transcriptional splicing efficiencies differ within genes and between cell types. *RNA*, 27:829–840, 2021. doi: 10.1261/rna.078662.120.

50. Simona Cugusi, Richard Mitter, Gavin P Kelly, Jane Walker, Zhong Han, Paola Pisano, Michael Wierer, Aengus Stewart, and Jesper Q Svejstrup. Heat shock induces premature transcript termination and reconfigures the human transcriptome. *Molecular Cell*, 82:1573–1588, 2022. doi: 10.1016/j.molcel.2022.01.007.

51. Yuka Nakazawa, Yuichiro Hara, Yasuyoshi Oka, Okiru Komine, Diana van den Heuvel, Chaowan Guo, Yasukazu Daigaku, Mayu Isono, Yuxi He, Mayuko Shimada, Kana Kato, Nan Jia, Satoru Hashimoto, Yuko Kotani, Yuka Miyoshi, Miyako Tanaka, Akira Sobue, Norisato Mitsutake, Takayoshi Suganami, Akio Masuda, Kinji Ohno, Shinichiro Nakada, Tomoji Mashimo, Koji Yamanaka, Martijn S Luijsterburg, and Tomoo Ogi. Ubiquitination of dna damage-stalled rnapii promotes transcription-coupled repair. *Cell*, 180:1228–1244, 2020. doi: 10.1016/j.cell.2020.02.010.

52. Ana Tufegdi Vidakovi, Richard Mitter, Gavin P Kelly, Michelle Neumann, Michelle Harreman, Marta Rodríguez-Martínez, Anna Herlihy, Juston C Weems, Stefan Boeing, Vesela Encheva, Liam Gaul, Laura Milligan, David Tollervey, Ronald C Conaway, Joan W Conaway, Ambrosius P Snijders, Aengus Stewart, and Jesper Q Svejstrup. Regulation of the rnapii pool is integral to the dna damage response. *Cell*, 180:1245–1261, 2020. doi: 10.1016/j.cell.2020.02.009.

53. Shahar Barbash and Thomas P. Sakmar. Length-dependent gene misexpression is associated with alzheimers disease progression. *Scientific Reports*, 7:190, 2017. doi: 10.1038/s41598-017-00250-4.

54. Akihide Takeuchi, Kei Iida, Toshiaki Tsubota, Motoyasu Hosokawa, Masatsugu Denawa, J B Brown, Kensuke Ninomiya, Mikako Ito, Hiroshi Kimura, Takaya Abe, Hiroshi Kiyonari, Kinji Ohno, and Masatoshi Hagiwara. Loss of sfpq causes long-gene transcriptopathy in the brain. *Cell Reports*, 23(5):1326–1341, 2018. doi: 10.1016/j.celrep.2018.03.141.

55. Brunilda Balliu, Matthew Durrant, Olivia de Goede, Nathan Abell, Xin Li, Boxiang Liu, Michael J. Gloudemans, Naomi L. Cook, Kevin S. Smith, David A. Knowles, Mauro Pala, Francesco Cucca, David Schlessinger, Siddhartha Jaiswal, Chiara Sabatti, Lars Lind, Erik Ingelsson, and Stephen B. Montgomery. Genetic regulation of gene expression and splicing during a 10-year period of human aging. *Genome Biol.*, 20:230, 2019. doi: 10.1186/s13059-019-1840-y.

56. Lorna W Harries, Dena Hernandez, William Henley, Andrew R Wood, Alice C Holly, Rachel M Bradley-Smith, Hanieh Yaghootkar, Ambarish Dutta, Anna Murray, Timothy M Frayling, Jack M Guralnik, Stefania Bandinelli, Andrew Singleton, Luigi Ferrucci, and David Melzer. Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing. *Aging Cell*, 10:868–878, 2011. doi: 10.1111/j.1474-9726.2011.00726.x.

57. Michael A Lodato, Rachel E Rodin, Craig L Bohrson, Michael E Coulter, Alison R Barton, Minseok Kwon, Maxwell A Sherman, Carl M Vitzthum, Lovelace J Luquette, Chandri N Yandava, Pengwei Yang, Thomas W Chittenden, Nicole E Hatem, Steven C Ryu, Mollie B Woodworth, Peter J Park, and Christopher A Walsh. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science*, 359:555–559, 2018. doi: 10.1126/science.aao4426.

58. Emily Mitchell, Michael Spencer Chapman, Nicholas Williams, Kevin J Dawson, Nicole Mende, Emily F Calderbank, Hyunchul Jung, Thomas Mitchell, Tim H H Coorens, David H Spencer, Heather Machado, Henry Lee-Six, Megan Davies, Daniel Hayler, Margarete A Fabre, Krishnaa Mahbubani, Federico Abascal, Alex Cagan, George S Vassiliou, Joanna Baxter, Inigo Martincorena, Michael R Stratton, David G Kent, Krishna Chatterjee, Kourosh Saeb Parsy, Anthony R Green, Jyoti Nangalia, Elisa Laurenti, and Peter J Campbell. Clonal dynamics of haematopoiesis across the human lifespan. *Nature*, 606:343–350, 2022. doi: 10.1038/s41586-022-04786-y.

59. Jiena Liu, Zhenzhen Wu, Jin He, and Yuming Wang. Cellular fractionation reveals transcriptome responses of human fibroblasts to uv-c irradiation. *Cell Death Dis.*, 13:177, 2022. doi: 10.1038/s41419-022-04634-x.

60. Gustavo Satoru Kajitani, Livia Luz de Souza Nascimento, Maira Rodrigues de Camargo Neves, Giovana da Silva Leandro, Camila Carrião Machado Garcia, and Carlos Frederico Martins Menck. Transcription blockage by dna damage in nucleotide excision repair-related neurological dysfunctions. *Semin Cell Dev Biol.*, 114:20–35, 2021. doi: 10.1016/j.semcdb.2020.10.009.

61. F. A. Wolf, Angerer P., and F. J. Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome Biology*, 2018. doi: https://doi.org/10.1186/s13059-017-1382-0.

62. M. Alex Ascensión, Olga Ibáñez-Solé, Iñaki Inza, Ander Izeta, and Marcos J Araúzo-Bravo. Triku: A feature selection method based on nearest neighbors for single-cell data. *GigaScience*, 11, 2022. doi: 10.1093/gigascience/giac017.

63. Leland McInnes, John Healy, Nathaniel Saul, and Lukas GroSSberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861.

64. V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9, 2019. doi: https://doi.org/10.1038/s41598-019-41695-z.

65. Olga Ibáñez-Solé, Alex M. Ascensión, Marcos J. Araúzo-Bravo, and Ander Izeta. Lack of evidence for increased transcriptional noise in aged tissues. *bioRxiv*, 2022. doi: 10.1101/2022.05.18.492432.

66. Simon Joost, Amit Zeisel, Tina Jacob, Xiaoyan Sun, Gioele La Manno, Peter Lönnerberg, Sten Linnarsson, and Maria Kasper. Single-cell transcriptomics reveals that differentiation and spatial signatures shape epidermal and hair follicle heterogeneity. *Cell Systems*, 3(3), 2016. doi: 10.1016/j.cels.2016.08.010.

67. Hubert W. Lilliefors. On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402, 1967. doi: 10.1080/01621459.1967.10482916.

Acknowledgements

We thank Irantzu Barrio for her support with the statistical analysis on gene length-dependent age-related transcriptional decay. We also thank Alex M. Ascensión, Javier Cabau-Laporta, Laura Yndriago, Sonia Alonso-Martin, Ander Matheu and David Otaegui for their thorough revision of the manuscript and for useful suggestions. This work was supported by grants from Instituto de Salud Carlos III (PI19/01621), cofunded by the European Union (European Regional Development Fund/ European Science Foundation, Investing in your future), Diputación Foral de Gipuzkoa. OI-S received the support of a fellowship from la Caixa Foundation (ID 100010434; code LCF/BQ/IN18/11660065), and from the European Union's Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No. 713673.

Author contributions

OI-S conceived and performed the experiments. AI conceived some experiments and supervised the work. Both authors wrote the manuscript.

Competing interests

The authors declare no competing interests.

Supplementary Figures

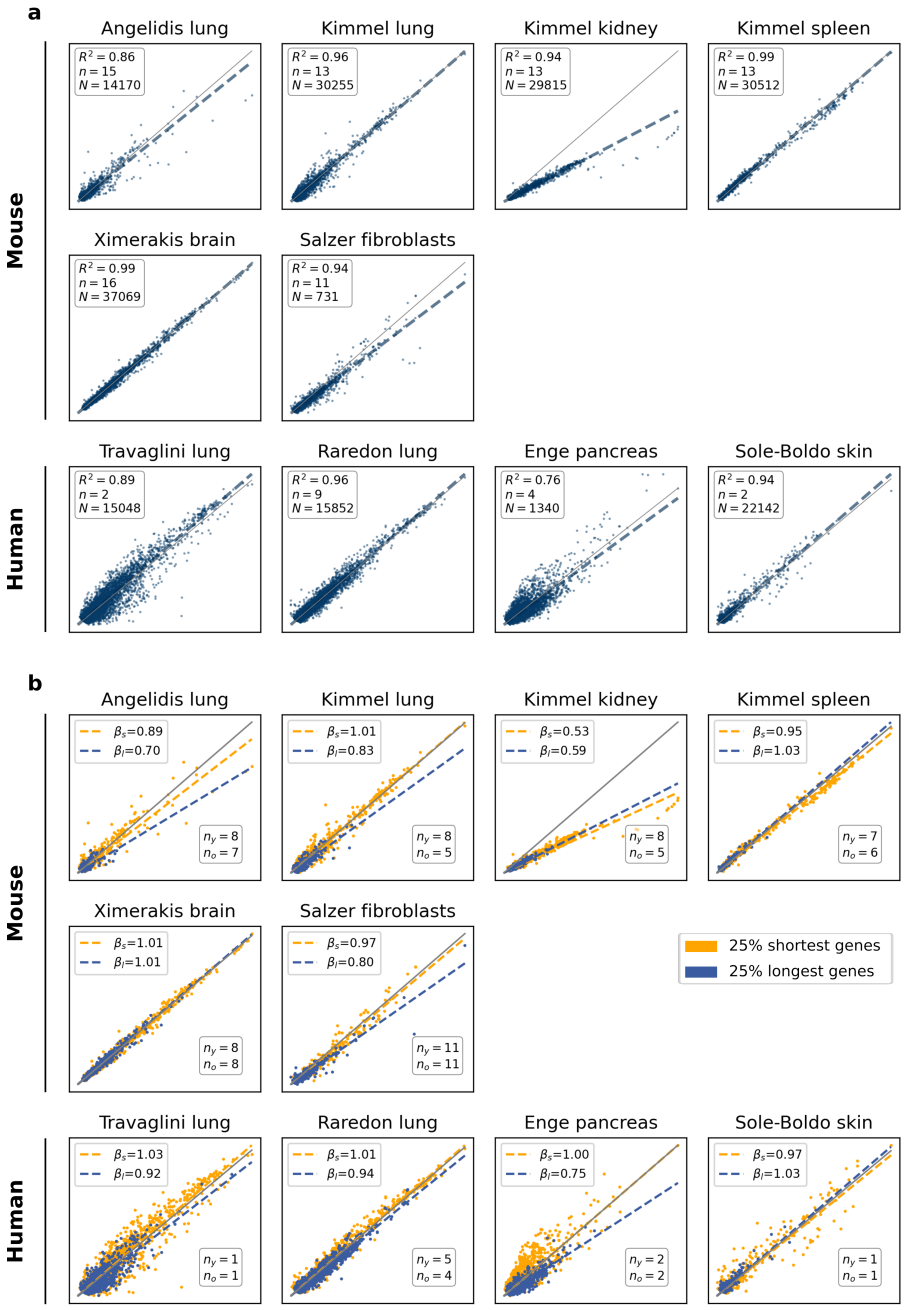


Figure S1. Downregulation of long genes with aging is replicated in several datasets of different species. a, Gene expression is conserved with aging in several datasets of different species. Average gene expression in old against young cells in six mouse and four human datasets of several tissues. R^2: coefficient of determination; N: total number of cells; n: number of biological replicates. **b**, Age-associated shutdown of transcription is found to be gene length-dependent in several datasets of different species. Slopes of the straight lines that fit the data for the 25% shortest (β_s) and the 25% longest genes (β_l). Number of biological replicates in each age category: young (n_y) and old (n_o).

Supplementary Tables

	25% shortest (Q1)			25% longest (Q4)		
	short_min	short_median	short_max	long_min	long_median	long_max
Bladder	65	5,373	10,218	58,647	104,278	2,270,723
Brain	63	5,402	10,143	52,003	93,157	1,211,426
Brain myeloid	69	5,521	10,418	57,559	103,285	2,257,271
Heart	69	5,196	9,713	52,057	95,055	2,270,723
Kidney	63	5,402	9,905	49,274	88,337	1,503,513
Liver	108	5,494	10,367	56,554	101,606	2,960,898
Lung	63	5,574	10,615	59,384	108,898	2,960,898
Muscle	64	5,897	11,150	63,517	118,298	2,960,898
Pancreas	67	5,526	10,471	55,760	100,580	2,960,898
Skin	63	5,411	10,151	56,744	101,808	2,960,898
Spleen	63	5,379	9,987	50,379	89,933	1,503,513
Thymus	63	5,538	10,287	54,303	99,058	2,960,898

Table S1. Length of the Q1 (25% shortest) and Q4 (25% longest) genes used in the analysis of Figure 1. The minimum, median and maximum gene lengths (bp) are shown for the two gene categories.

	Q1		Q1-Q2		Q1-Q3		Q1-Q4	
	Est. (SE)	p-val	Est. (SE)	p-val	Est. (SE)	p-val	Est. (SE)	p-val
Bladder	1.02 (<0.01)	0	-0.03 (0.01)	<0.001	-0.05 (0.01)	<0.001	-0.07 (0.01)	<0.001
Brain	0.86 (<0.01)	0	-0.25 (0.01)	<0.001	-0.35 (0.01)	<0.001	-0.43 (0.01)	<0.001
Brain myeloid	1.03 (<0.01)	0	-0.06 (0.01)	<0.001	-0.11 (0.01)	<0.001	-0.29 (0.01)	<0.001
Heart	1.02 (<0.01)	0	-0.09 (0.01)	<0.001	-0.17 (0.01)	<0.001	-0.28 (0.01)	<0.001
Kidney	0.93 (<0.01)	0	-0.03 (0.01)	3.85E-02	-0.08 (0.02)	<0.001	-0.19 (0.02)	<0.001
Liver	0.86 (<0.01)	0	-0.06 (0.01)	<0.001	-0.08 (0.01)	<0.001	-0.20 (0.02)	<0.001
Lung	1.16 (0.01)	0	-0.24 (0.02)	<0.001	-0.39 (0.02)	<0.001	-0.50 (0.02)	<0.001
Muscle	1.26 (0.01)	0	-0.36 (0.02)	<0.001	-0.50 (0.02)	<0.001	-0.62 (0.02)	<0.001
Pancreas	0.85 (<0.01)	0	-0.03 (0.01)	2.83E-02	-0.06 (0.01)	<0.001	-0.14 (0.01)	<0.001
Skin	1.09 (<0.01)	0	-0.13 (0.01)	<0.001	-0.21 (0.01)	<0.001	-0.29 (0.01)	<0.001
Spleen	0.99 (<0.01)	0	0.05 (0.01)	<0.001	-0.03 (0.01)	3.80E-02	-0.19 (0.01)	<0.001
Thymus	1.02 (<0.01)	0	-0.13 (0.02)	<0.001	-0.25 (0.02)	<0.001	-0.41 (0.02)	<0.001

Table S2. Linear models fit on short and long genes are significantly different in 12 murine aging mouse datasets. We test for the difference between the slope that best fits the old *vs* young average gene expression using the Q1 genes (25% shortest) and the slope that corresponds to each of the other three quartiles (Q2, Q3, Q4). Q1-Q2, Q1-Q3 and Q1-Q4 represent the differences between the slopes fitted on Q1 and each of the quartiles. Est. (estimate), SE (standard error), p-val (p-value).

	U statistic	p-value
Bladder	28948.5	5.23e-10
Brain	27401.0	2.52e-10
Brain myeloid	13075.0	1.45e-43
Heart	12005.5	5.54e-45
Kidney	22024.0	9.91e-21
Liver	31636.0	0.000227
Lung	10844.0	7.16e-52
Muscle	8774.5	7.31e-59
Pancreas	25380.0	6.00e-12
Skin	12953.5	1.35e-44
Spleen	19386.0	5.45e-25
Thymus	10888.5	1.97e-50

Table S3. Mann-Whitney test comparing lengths of DEG between young and old cells U statistic and p-value associated with each comparison. The test compares the mean log_{10} gene length (bp) of the top 300 DEGs between young and old cells in 12 murine tissues (shown in Figure 2).

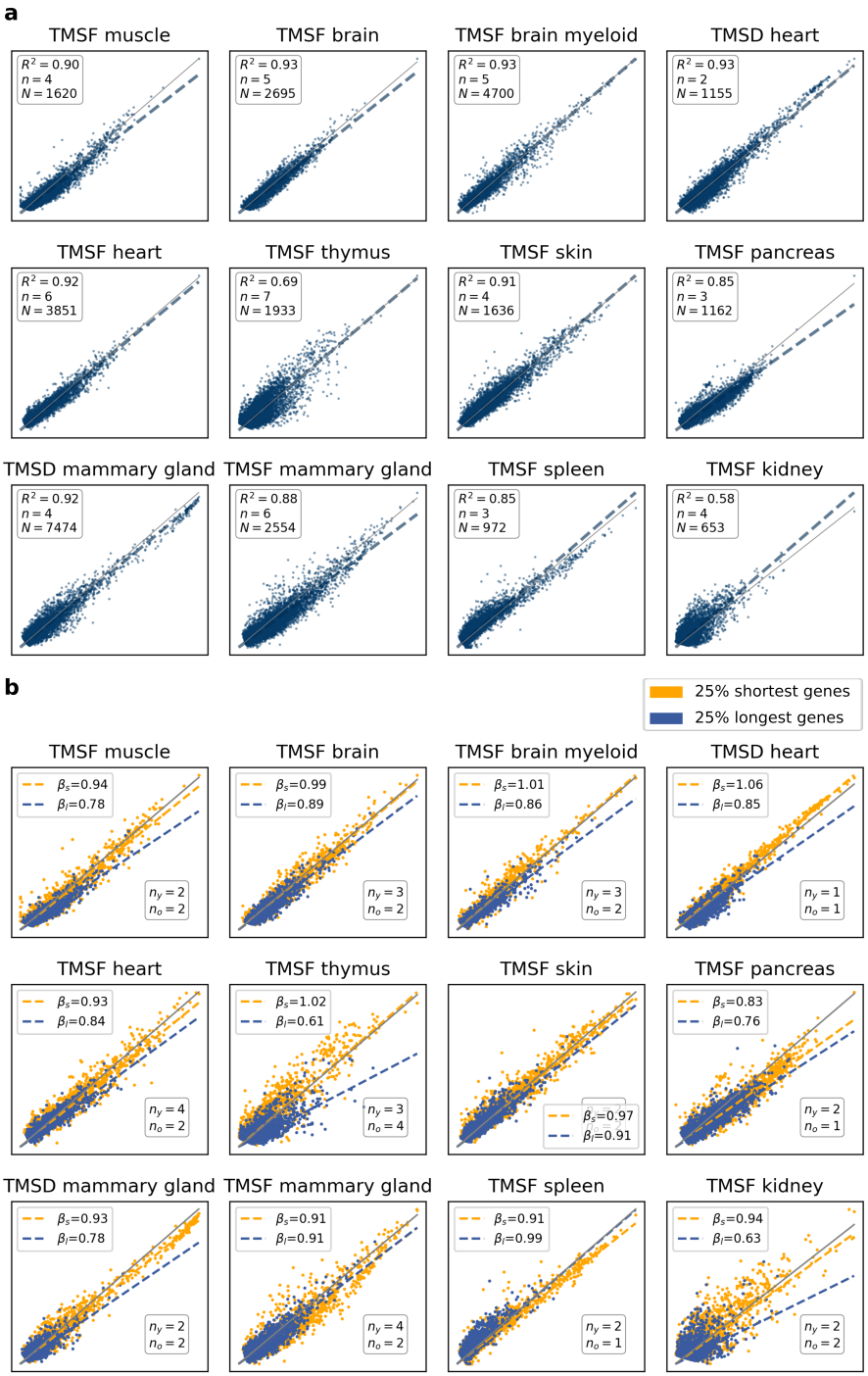


Figure S2. Age-associated shutdown of transcription is also detected in 18 month-old females. a, Gene expression is highly conserved but shows a detectable decay with aging in 18 month old female mice as well. Scatter plots showing the average gene expression in 18-month old female mice against average gene expression in 3 month-old female mice in 12 tissues from the *TMS FACS* and the *TMS droplet* datasets [17]. Each dot represents a gene. N: number of single cells; n: number of biological replicates. R^2: coefficient of determination. **b**, Age-associated shutdown of transcription preferentially affects long genes. The scatter plots show the average gene expression in 18 month-old versus in 3 month-old female mice. The top 25% and bottom 25% of the total genes according to their gene length are shown in blue and yellow, respectively. β_s and β_l represent the slopes of the straight lines that best fit the data points corresponding to *short* and *long* genes, respectively. Number of young (n_y) and old (n_o) biological replicates.

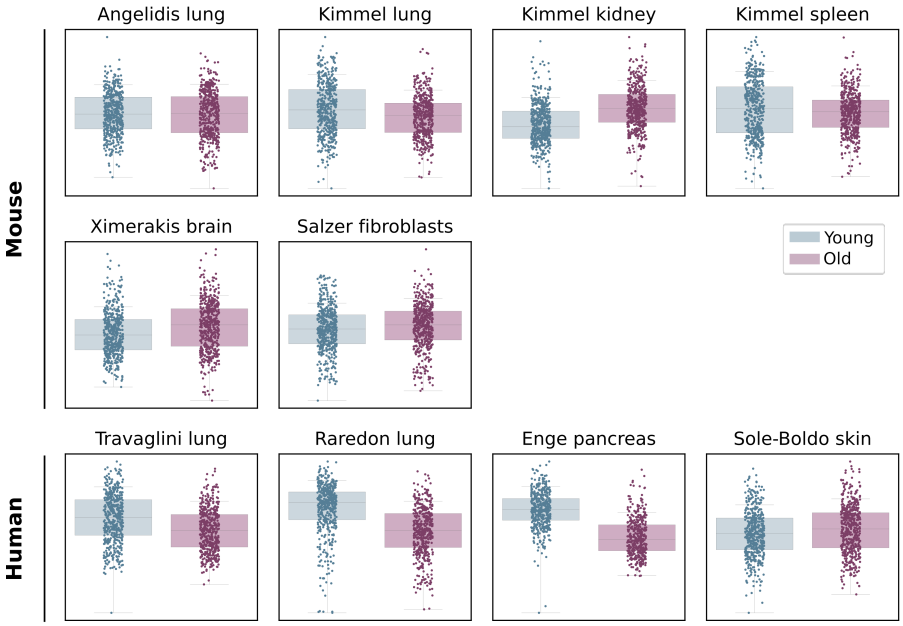


Figure S3. Downregulation of long genes is found in several datasets of different species. Top 300 DEGs between young and old cells in 10 independent aging datasets from mouse and human. The 300 differentially expressed genes between young and old individuals were obtained using the Wilcoxon method.

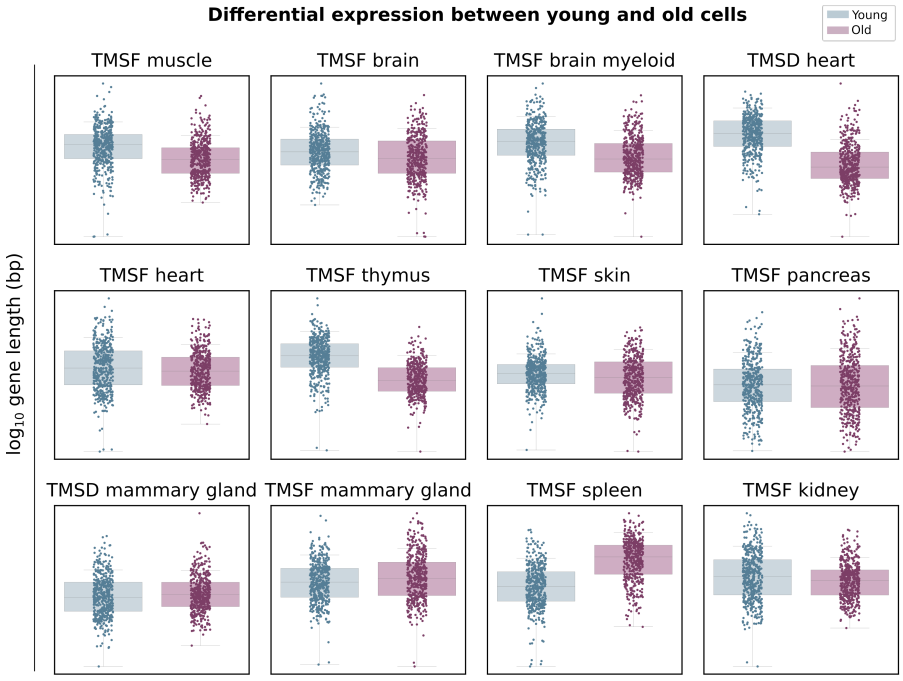


Figure S4. Downregulation of long genes is also detected in 18 month-old females. Top 300 DEGs between young and old cells in 12 aging datasets from the Tabula Muris Senis. The 300 differentially expressed genes between 3 months old and 18 months old female mice were obtained using the Wilcoxon method.

		Q1		Q1-Q2		Q1-Q4		Q1-Q4	
		Est. (SE)	**p-val**	**Est. (SE)**	**p-val**	**Est. (SE)**	**p-val**	**Est. (SE)**	**p-val**
Lin	H vs UV	0.91 (<0.01)	0	-0.07 (<0.01)	<0.001	-0.10 (0.01)	<0.001	-0.11 (0.01)	<0.001
	VD vs UV	0.91 (<0.01)	0	-0.08 (<0.01)	<0.001	-0.10 (0.01)	<0.001	-0.13 (0.01)	<0.001
	H vs VD	0.98 (<0.01)	0	0.01 (<0.01)	0.392	-0.01 (0.01)	0.201	0.00 (0.01)	0.903
Goldfar.	H vs Smoker	1.01 (<0.01)	0	-0.03 (<0.01)	<0.001	-0.06 (0.01)	<0.001	-0.12 (0.01)	<0.001
	WT vs ADH5KO	1.00 (<0.01)	0	0.02 (<0.01)	<0.001	0.01 (0.01)	0.171	0.01 (0.01)	0.0969
Mulder.	WT vs CSBKO	1.00 (<0.01)	0	0.03 (<0.01)	<0.001	0.03 (<0.01)	<0.001	0.04 (0.01)	<0.001
	WT vs DKO	0.99 (<0.01)	0	-0.02 (<0.01)	4.68E-03	-0.05 (0.01)	<0.001	-0.06 (0.01)	<0.001
	GC: ct vs UV	0.94 (<0.01)	0	0.00 (<0.01)	0.691	0.01 (0.01)	9.30E-03	0.00 (0.01)	0.893
Wang	mut: ct vs UV	0.94 (<0.01)	0	-0.04 (<0.01)	<0.001	-0.07 (0.01)	<0.001	-0.10 (0.01)	<0.001
	ct: GC vs mut	0.97 (<0.01)	0	0.02 (<0.01)	<0.001	0.02 (<0.01)	<0.001	0.00 (<0.01)	0.958
	UV: GC vs mut	0.99 (<0.01)	0	-0.03 (<0.01)	<0.001	-0.08 (<0.01)	<0.001	-0.12 (0.01)	<0.001

Table S4. Statistical significance of the analyses done on premature aging datasets.

	Q1		Q1-Q2		Q1-Q3		Q1-Q3	
	Est. (SE)	**p-val**	**Est. (SE)**	**p-val**	**Est. (SE)**	**p-val**	**Est. (SE)**	**p-val**
TMSD F (3-18)	1.06 (0.00)	<0.001	-0.07 (0.01)	<0.001	-0.14 (0.01)	<0.001	-0.21 (0.01)	<0.001
TMSD F (3-21)	1.06 (0.00)	<0.001	-0.04 (0.01)	<0.001	-0.12 (0.01)	<0.001	-0.22 (0.01)	<0.001
TMSD M (1-18)	0.97 (0.00)	<0.001	-0.04 (0.01)	<0.001	-0.05 (0.01)	<0.001	-0.09 (0.01)	<0.001
TMSD M (1-24)	0.98 (0.00)	<0.001	-0.03 (0.01)	<0.001	-0.04 (0.01)	<0.001	-0.07 (0.01)	<0.001
TMSF F (3-18)	0.93 (0.00)	<0.001	-0.01 (0.01)	7.84E-2	-0.05 (0.01)	<0.001	-0.09 (0.01)	<0.001
TMSF M (3-24)	1.02 (0.01)	<0.001	-0.09 (0.01)	<0.001	-0.17 (0.01)	<0.001	-0.28 (0.01)	<0.001

Table S5. Output of the statistical analysis comparing the effects of the different gene length groups based on a linear model with interaction. We test for the difference between the slope that best fits the old *vs* young average gene expression using the Q1 genes (25% shortest) and the slope that corresponds to each of the other three quartiles (Q2, Q3, Q4). Q1-Q2, Q1-Q3 and Q1-Q4 represent the differences between the slopes fitted on Q1 and each of the quartiles.