

Aging associated changes of transcriptional elongation speed and transcriptional error rate



Inaugural-Dissertation
zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Antonios Papadakis
geboren in Heraklion

Köln, February 2024

Berichtersteller:

Prof. Dr. Andreas Beyer

Dr. Peter Tessarz

Tag der mündlichen Prüfung: 13/02/2023

Table of Contents

Acknowledgments	i
Erklärung zur Dissertation	iii
Abstract	iv
Chapter 1. Introduction	1
1.1 Overview	1
1.2 Transcription	3
1.2.1 Transcriptional initiation	4
1.2.2 Transcriptional elongation	4
1.2.3 Transcriptional termination	5
1.3 RNA Pol II elongation speed	6
1.3.1 Methods used to estimate elongation speed in eukaryotes	6
Biochemical approaches in one or few genes	6
Imaging approaches in one or few genes	6
Genome-wide approaches	6
1.3.2 Pol II speed regulation	9
1.3.3 Processes regulated by Pol II speed	11
Splicing	11
Transcriptional initiation and termination	11
Regulation of chromatin structure and gene expression	12
Transcriptional fidelity	12
1.4 Transcription elongation errors	13
1.4.1 Transcription errors and correction mechanisms	13
1.4.2 Physiological consequences of transcription elongation errors	13
1.4.3 Measurements of transcription errors	14
1.5 Single-cell RNA sequencing and unique molecular identifiers	17
1.5.1 Overview of commonly used scRNAseq methods	17
1.5.2 Unique molecular identifiers	18
1.6 Aims of the Project	19
Chapter 2. The effects of aging-associated changes in transcriptional elongation on metazoan longevity	20
2.1 Introduction	20
2.2 Results	21
2.3 Materials and methods	31
2.3.1 Biological materials	31

2.3.2 Biochemistry and molecular biology methods	33
2.3.3 Computational methods	37
2.5 Contributions	42
Chapter 3. Age-related changes in transcriptional fidelity across tissues	43
3.1 Introduction	43
3.2 Results	45
3.3 Materials and methods	51
3.4 Discussion	54
3.5 Contributions	55
Chapter 4. General discussion	56
4.1 The effects of aging-associated changes in transcriptional elongation on metazoan longevity	56
4.1.1 Limitations	56
4.1.2 Future directions	57
4.2 Age-related changes in transcriptional fidelity across tissues	58
4.2.1 Limitations	59
4.2.2 Future directions	59
Appendix A: Supplementary Figures	60
Appendix B: Supplementary Tables	77
Bibliography	80

Acknowledgments

This thesis has a single author, but it was a collaborative effort, with many individuals contributing directly and indirectly to the final product.

First, I am deeply grateful to Andreas Beyer, my supervisor, who welcomed me into his research group all the way back in 2017 and has consistently demonstrated enthusiasm, attentiveness, and, most of all, understanding throughout the process. His mentorship and support have been invaluable to me, and not only would I not have been able to complete this thesis without him, but I would also be a worse scientist overall.

The Beyer research group was integral to the success of this project, and I am especially thankful to Cedric Debès, who began the transcription elongation rate studies that inspired the project in the first place and provided guidance and mentorship during my first challenging year. I also want to express my gratitude to my student supervisees: Jonatan Gabre and Isabell Brusius. Jonatan played an important role in the development of the methods explored within Chapter 2 of the thesis and Isabell provided crucial help with the splicing analysis.

As a bioinformatician, I would not be able to do any analysis without collaborators who got their hands dirty generating the data. The Nephrolab in CECAD and the groups of Adam Antebi and Linda Partridge from the Max Planck Institute for Biology of Ageing provided vital contributions for Chapter 2 of the thesis. A special thanks needs to go to the entire group of Akis Papantonis, now in Göttingen, who helped tremendously with both chapters of the thesis, both experimentally and conceptually. I spent hours discussing the project with Akis himself, who was always there to help and provide unwavering support. It is almost certain that my scientific career would be completely different if not for him. An additional word of appreciation goes to two members of Akis' group. Konstantinos Sofiadis, an old friend from Crete, helped my integration to Germany immensely and Natasa Josipovic who conducted very important experiments for my work and was supportive during my first years here.

Beyond the help they provided me with their research and ideas, the Beyer group was also the friendliest and most pleasant workplace environment I have ever worked in. Although most of the members have changed since I first arrived, as a whole they have helped me and inspired me during my years here. A special word of gratitude goes to Luise Nagel, who proofread parts of this thesis and massively improved it, Tim Padvitski, who has shown friendship and support in ways that would require a whole new thesis to describe, and Jan Grossbach, with whom I have shared some of the most interesting conversations during our long coffee breaks and who has always stood by me when it mattered. Without them, and without Jonatan, Fabian, Carolina, Paula and many other past and present members of the group, my scientific journey would be much poorer. I also have to thank our secretary, Kay Heitplatz, whose dedication in helping me navigate countless bureaucratic challenges, even when they were mostly my fault, has always impressed me.

My sincere thanks also to my fellow workers and students here in CECAD, especially Dimitrije Stankovic, Akos Gyenis and Péter Szántó, for providing both advice and comradeship

during my time here. Thanks also to my collaborators and assistants in all the other projects I have been working on during the lengthy process of my PhD and thanks to my fellow students at the MPI-AGE and CMMC. Thanks to all the other people I have inexcusably failed to mention here because the acknowledgments should be shorter than the discussion section of the thesis.


My family, of course, has been an endless wellspring of love and support. I thank them for all the help they have provided over the years. Finally, my deepest and sincerest thanks to my wife, Virna, without whose presence this very long journey would be a great deal less interesting and fun.

Erklärung zur Dissertation

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel und Literatur angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Werken dem Wortlaut oder dem Sinn nach entnommen wurden, sind als solche kenntlich gemacht. Ich versichere an Eides statt, dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie - abgesehen von unten angegebenen Teilpublikationen und eingebundenen Artikeln und Manuskripten - noch nicht veröffentlicht worden ist sowie, dass ich eine Veröffentlichung der Dissertation vor Abschluss der Promotion nicht ohne Genehmigung des Promotionsausschusses vornehmen werde. Die Bestimmungen dieser Ordnung sind mir bekannt. Darüber hinaus erkläre ich hiermit, dass ich die Ordnung zur Sicherung guter wissenschaftlicher Praxis und zum Umgang mit wissenschaftlichem Fehlverhalten der Universität zu Köln gelesen und sie bei der Durchführung der Dissertation zugrundeliegenden Arbeiten und der schriftlich verfassten Dissertation beachtet habe und verpflichte mich hiermit, die dort genannten Vorgaben bei allen wissenschaftlichen Tätigkeiten zu beachten und umzusetzen. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Teilpublikationen:

Debès, C. et al. Aging-associated changes in transcriptional elongation influence metazoan longevity. Preprint at <https://doi.org/10.1101/719864> (2022).

19.12.2022, Antonios Papadakis, 

Datum, Name und Unterschrift

Abstract

Aging organisms show a pervasive decline in cellular function, with important implications for healthspan and lifespan. Aging associated impairment of gene transcription from Pol II are believed to underlie a large part of this phenotype. Elongation is a particularly important step of transcription since it regulates a lot of cotranscriptional processes, however the exact molecular mechanisms involved in its changes with aging remain unclear.

In this thesis, we report the effects of aging in various transcriptional processes across different eukaryotes. We use a combination of previously published and newly-generated next generation sequencing data to understand the mechanisms of aging associated changes in Pol II speed and fidelity

We profiled and analyzed genome-wide, aging-related changes in transcriptional processes across different organisms: nematode worms, fruit flies, mice, rats and humans. The average transcriptional elongation speed (Pol II speed) increased with age in all five species. Along with these changes in elongation speed, we observed changes in co-transcriptional processes that are partially regulated by elongation, including splicing alterations, the formation of more circular RNAs and loss of transcriptional fidelity. Two lifespan-extending interventions, dietary restriction and lowered insulin/Igf signaling, both partially reversed some of these aging-related changes. Remarkably, genetic variants of Pol II that reduced its speed in worms and flies increased their lifespan, which proves the importance of elongation rate for organismal longevity. Similarly, reducing Pol II speed by overexpressing histone components, to counter age-associated changes in nucleosome positioning, also extended lifespan in flies and the division potential of human cells. Our findings uncover fundamental molecular mechanisms underlying animal aging and lifespan-extending interventions, and point to possible preventative measures.

Furthermore, we developed a new computational pipeline, `scErrorRate`, that utilizes UMI-based single-cell data to estimate transcriptional error rate. It is a computational approach that does not require the onerous process of rolling circle-based technologies. Using `scErrorRate`, we were able to profile the error spectrum of Pol II in mice and human cell culture. For the first time, we characterized changes in transcriptional fidelity caused by aging and senescence, showing an overall increase in transcriptional misincorporations. Taken together, this work provides new insight on the fundamental molecular mechanisms underlying aging.

Chapter 1. Introduction

1.1 Overview

Every living being will experience the end of life. For many of us, the ultimate cause of this irreversible erasure will be aging, a gradual process of organismal deterioration that occurs with time. Aging impairs a wide range of cellular processes, leading to a decline in cellular fitness and progressive loss of function. It is an important risk factor for various chronic diseases such as cancer, cardiovascular and metabolic pathologies, and neurodegeneration¹. Considering that there has been a significant global increase in life expectancy globally, this means that aging-related diseases are an increasingly greater socioeconomic burden². As a consequence, increasing “healthspan” and slowing down age-related disorders is vital³. Understanding the molecular mechanisms at work is necessary if we ever hope to discover preventative measures and for the future of geroscience in general, but it remains a work in progress.

On the molecular and cellular level, there are multiple hallmarks that characterize aging⁴, either as responses to age-associated deterioration or as causes of cell dysfunction. They include genomic instability, loss of proteostasis, deregulation of nutrient sensing, altered intercellular communication, mitochondrial dysfunction, telomere attrition, stem cell exhaustion, altered intercellular communication and epigenetic alterations. One of the most important processes affected by aging that is related to these hallmarks is transcription, the first step of gene expression. During aging, transcriptome composition in animals changes dramatically⁴. Furthermore, aging causes a significant increase in variability and errors in the expression of genes^{5,6}.

The process of transcription and its regulation have been studied extensively with a great variety of specialized biochemical and imaging approaches, in various contexts and organisms. Initially, studies of transcriptional regulation mostly focused on its first step, transcriptional initiation. However, it has become increasingly clear that the other stages of transcription are critical for the control of gene expression, with transcriptional elongation being of particular interest. During elongation, the enzyme RNA polymerase II travels the entire length of the gene in a step-wise, nucleotide-by-nucleotide process. This leads to the synthesis of an RNA chain complementary to the template strand of the DNA.

Elongation is very important for correct mRNA production, since it controls multiple cotranscriptional pre-mRNA processing steps, like splicing, polyadenylation and termination^{7,8}. Correct regulation of these events is vital, since errors in these steps can result in severe decline of cellular health. Given the fact that aging causes significant changes in the output of RNA synthesis, it is reasonable to assume that it could affect elongation itself. However, relatively little attention has focused on transcriptional elongation in aging studies. Much remains to be discovered about the effects of aging on the kinetics and fidelity of transcription and the impact of these changes on aging-associated decline of function.

The speed of RNA polymerase II is an important determinant of the composition of the transcriptome. It regulates fundamental co-transcriptional processes, like transcriptional termination and splicing, controlling the production of alternative transcript isoforms⁹ and circular RNAs^{7,10}. However, we lack full understanding of how Pol II speed is affected by aging. Similarly, there is a lack of published research about the impact of aging on transcriptional fidelity^{5,11}. Transcription errors can mimic detrimental mutations^{12,13} and apply pressure to protein quality control mechanisms⁵. Analyzing the dynamics of Pol II speed and error rate could provide insight in aging-related disease progression and deterioration of physiological processes.

However, estimating Pol II speed and Pol II fidelity in living organisms requires elaborate biochemical approaches that are time-consuming and onerous. One of the main driving impulses of this thesis is an effort to overcome these methodological limitations that have inhibited further *in vivo* elucidation of the kinetics and mechanisms of Pol II elongation.

1.2 Transcription

Identical genomes produce a great variety of cellular phenotypes in the same organism. Most of this variation can be attributed to differences in gene expression, a complex array of mechanisms by which proteins are synthesized using information encoded in the DNA. The rate of protein synthesis is regulated at multiple steps during the process of gene expression, but the primary one is the very first step: transcription.

The vast variety of cellular phenotypes that are produced by identical genomes can mostly be attributed to differences in gene expression, a complex array of mechanisms by which proteins are synthesized using information encoded in the DNA. The rate of protein synthesis is regulated at multiple steps during the process of gene expression, but the primary one is the very first step: transcription.

Transcription is the synthesis of an RNA molecule from DNA, creating a copy of the information contained in a gene. The resulting RNA molecule is called **messenger RNA (mRNA)** and it contains the necessary information to construct a protein¹⁴. The sequence of the transcript is complementary to the DNA sequence of the genes, with the difference that thymines are replaced by uracils. Transcription is a precisely regulated process that allows both the maintenance of cellular homeostasis and the adjustment of the composition of the cellular proteome in response to environmental cues and various external stimuli¹⁵.

RNA polymerase is the main protein responsible for transcription. While there is only one RNA polymerase in prokaryotes and archaea, the transcription in eukaryotes is performed by five different polymerases. Two of them (Pol IV and Pol V) are exclusive to plants. The other three polymerase enzymes, Pol I, II and III¹⁶, exist in all eukaryotes. Even though the general process of the transcription cycle is almost identical for all RNA polymerases, the transcription factors that associate with the polymerases and the regions of the genome that they transcribe are very different. Pol I catalyzes the synthesis of the precursor ribosomal RNAs (pre-rRNAs) which are then processed into mature rRNAs, integral parts of the cellular ribosome. Pol III transcribes transfer RNAs, the 5S rRNA and many small RNAs. Pol II is responsible for the transcription of all protein-coding regions of the genome and multiple non-coding ones (including snRNAs, snoRNAs and miRNAs).

In yeast, replicative aging has a remarkable impact on Pol II transcription, leading to global upregulation of Pol II gene expression¹⁷. In higher eukaryotes, the effect is milder, with less than 5% of the genes showing age-associated differential expression¹⁸. A lot of these genes have important age-related functions. For instance, the expression of Pol II genes in stress response and inflammation pathways is commonly induced with aging^{19,20}. Metabolic and DNA repair genes are commonly downregulated²¹. These pathways play an important role in aging; thus, it is worth exploring how their transcription functions in greater detail.

Pol II is a protein complex which consists of 12 subunits and which is conserved throughout eukaryotes. It doesn't mediate transcription on its own; a very wide group of proteins called

transcription factors (TFs) can either activate or repress transcription by binding on the DNA, on the polymerase itself or on other proteins¹⁴.

Transcription by Pol II proceeds in the following three general steps:

1) **Transcriptional initiation.** The polymerase binds to the DNA of the gene at a region called the **promoter**. The pre-initiation protein complex is assembled. It opens the DNA and the polymerase starts transcription.

2) **Transcriptional elongation.** The polymerase moves along one strand of DNA in the 3' to 5' direction. For each nucleotide in the DNA template, RNA polymerase adds a matching (complementary) RNA nucleotide to the 3' end of the RNA strand.

3) **Transcriptional termination.** Pol II transcribes a DNA sequence that signals cleavage of the 3' end of the RNA molecule. Transcription ends and the RNA transcript is released along with the polymerase.

1.2.1 Transcriptional initiation

Transcription by RNA Pol-II first requires the assembly of a pre-initiation complex (PIC) bound at the promoter. The minimal PIC is composed of Pol-II, the Mediator complex and six general transcription factors: TFIIA, TFIIB, TFIID, TFIIIE, TFIIF and TFIIH²². Regulatory components that render the promoter accessible, like chromatin remodelers and histone acetyltransferases, are also often involved in the process²³. Once the PIC is assembled on the promoter, Pol II escapes the promoter and RNA synthesis commences.

After elongating 25-50 bp, RNA Pol II pauses. This state is called promoter-proximal pausing²⁴ and it is one of the key rate-limiting steps in the production of RNA²⁵. The paused Pol II is stabilized by two factors, the DRB-sensitivity inducing factor (DSIF)²⁶ and the negative elongation factor (NELF)²⁷. The positive transcription elongation factor b (P-TEFb) releases Pol II from this pausing by phosphorylating multiple proteins, including the CTD of Pol II, NELF and the DSIF subunit Spt5²⁸. NELF dissociates from Pol II and DSIF switches from a repressing factor to an activating one.

1.2.2 Transcriptional elongation

After pause release, RNA Pol II begins the process of productive transcription elongation. The transcription elongation complex is minimally composed of three parts: the double-stranded DNA template, the nascent RNA that is being synthesized and Pol II. Elongation occurs in steps. The complex contains a DNA-RNA duplex known as a transcription bubble²⁹. Because of the bubble, the most 3' end of RNA is positioned at the active site of RNA polymerase. The incoming nucleotide binds to the active site based on its complementarity to the next base on the DNA template. Subsequently, the formation of a phosphodiester bond between the 3'-OH group of the nascent RNA and the new nucleotide is catalyzed. Finally, the polymerase moves to the next template position³⁰. Overall, Pol II elongation proceeds in the 5'-3' direction in steps, but it is not a monotonous process, as the polymerase can be interrupted by pauses, premature disengagements

and even backtracking³¹. This has the potential to cause premature termination, creating transcripts that become non-coding RNAs, new proteins or targets for rapid degradation³².

A chromatin structure that facilitates the passage of Pol II through the gene while preventing intragenic transcription is vital for elongation³³. This requires the recruitment of multiple histone chaperones and elongation factors. Two of these very important proteins for rapid elongation are PAF (Polymerase Associated Factor)³⁴ and SPT6³⁵. Along with DSIF, they associate with Pol II to create an activated transcription elongation complex. PAF strongly allosterically stimulates the polymerase³⁶ and recruits histone chaperones (FACT³⁷), histone modifiers (BRE1³⁸, DOT1L³⁹) and histone remodellers⁴⁰ (CHMD1). The chaperones SPT6 and FACT are required for maintaining the proper chromatin structure during elongation⁴¹⁻⁴³.

1.2.3 Transcriptional termination

At the end of almost all eukaryotic protein-coding genes, there is a polyadenylation signal (PAS)⁴⁴. The PAS contains an AAUAAA motif and it is recognized by the cleavage and polyadenylation complex (CPA), which is composed of multiple factors including the cleavage and polyadenylation specificity factor (CPSF), the cleavage stimulatory factor and the cleavage factors I and IIm subcomplexes. The association of CPA with PAS leads to the cleavage of the pre-mRNA, which is then polyadenylated at the cleaved 3' end^{45,46}.

There are two main models that explain how transcription terminates after PAS transcription:

- 1) The allosteric model⁴⁷, according to which transcription of the polyadenylation site induces conformational changes to the complex, mediated by the binding of termination factors or loss of elongation factors.
- 2) The "torpedo" model⁴⁸, according to which the cleavage by the CPA creates a new, unprotected 5' end on the nascent RNA, allowing the entry of the 5'-3' exoribonuclease 2 (Xrn2 in humans, Rat1 in yeast). The exonuclease degrades the nascent RNA synthesized beyond the termination signal and chases the Pol II, dislodging it from the DNA once it catches up to it⁴⁹.

The two models are not mutually exclusive. A unified model has recently been proposed in which the dephosphorylation of the DSIF subunit Spt5 slows down elongation, allowing the Xrn2 exoribonuclease to catch up with Pol II and trigger transcript release⁵⁰.

After termination, the newly-cleaved nascent RNA undergoes several post-transcriptional modifications while the polymerase is again available for promoter binding.

1.3 RNA Pol II elongation speed

Until the last couple of decades, the focus of studies about the transcriptional regulation of gene expression was transcriptional initiation. Elongation was not considered as important for expression regulation. New research, however, has indicated that the elongation speed is highly dynamic and tightly coupled with other cellular processes that define the composition of the transcriptome.

In the first part of this section, the existing technologies for Pol II speed estimation will be reviewed and an evaluation of their advantages and disadvantages. In the second part, the existing literature about the known factors that determine Pol II speed will be reviewed. The third and final part concerns our current knowledge of the effects of Pol II speed on many cellular processes.

1.3.1 Methods used to estimate elongation speed in eukaryotes

Biochemical approaches in one or few genes

Early experiments used radioisotope pulse labeling of RNA to measure transcription speed in HeLa cells, finding a rate of 3-6 kb/min⁵¹. In the first half of the nineties, elongation rates were measured with in situ hybridization in the *Drosophila* Ubx gene⁵² and with nuclear run-on assays in the *Drosophila* heat shock genes⁵³ and the human dystrophin gene⁵⁴. These methods yielded estimates ranging from 1.1 to 2.5 kb/min. Chromatin immunoprecipitation of Pol-II followed by PCR has also been used to estimate elongation rates (1-2 kb/min)⁵⁵. More refined quantitative approaches like tiling arrays⁵⁶ and RT-PCR⁵⁷ gave somewhat higher speed estimates (3.1-3.8 kb/min). While these approaches give a first indication of the transcriptional speed, they are limited to one or few genes and therefore do not provide a global image of transcription.

Imaging approaches in one or few genes

Fluorescence recovery after photobleaching (FRAP) is a microscopy-based approach used to investigate molecular dynamics *in vivo*. It can be used to monitor Pol II elongation during steady state transcription. FRAP experiments for Pol II speed estimation use MS2 proteins fused with GFP that bind to nascent RNA transcripts containing a series of repeated MS2 loops⁵⁸. The speed estimated with this method ranged from 0.3 to 4.3 kb/min⁵⁸⁻⁶². In one MS2 experiment, transcription from HIV promoters was shown to reach up to 50-100 kb/min⁶³. Labs attempting dual fluorescence detection of nascent transcripts in yeast⁶⁴ and *Drosophila*⁶⁵ reported estimates ranging from 0.88 to 3.66 kb/min. Measurements with FRAP experiments using fused Pol II in *Drosophila* were also tried, yielding a range of speeds between 1.1 to 1.5 kb/min^{66,67}. These approaches also suffer from the fact that they can be applied to a limited number of genes.

Genome-wide approaches

During the last few years, the rapid development of next-generation sequencing (NGS) methods has significantly reduced the price of genome-wide tracking of nascent transcription. There are currently multiple approaches utilizing NGS in the context of investigating Pol II speed, which can be divided in two broad categories:

1) Methods measuring Pol II distribution on chromatin

Chromatin immunoprecipitation is used to investigate interactions between proteins and the DNA *in vivo*. The proteins of interest are crosslinked with the DNA and then the complexes are immunoprecipitated via protein-specific antibodies. The DNA is then purified from the complexes and the specific nucleotide sequences that interact with the proteins can be identified with PCR, qPCR, microarrays (ChIP-on-chip) or sequencing (ChIP-Seq)⁶⁸. Using this method, Pol-II binding can be mapped genome-wide, providing a snapshot of the Pol II distribution throughout the genome. Chip-Seq cannot be used for estimation of elongation speed since it is very highly enriched in stalled and paused polymerase molecules; it is impossible to know whether the polymerase was elongating at the time.

Native elongating transcript sequencing (NET-seq) can provide genome-wide mapping of Pol II density at nucleotide resolution by identifying the 3' ends of nascent RNA⁶⁹. Chromatin-bound RNA can be extracted from the polymerase and sequenced. The technique was successfully used first in yeast and then in bacteria⁷⁰, identifying multiple new pause sites through the genome and increasing our understanding of the pervasiveness of backtracking. With mammalian native elongating transcript sequencing (mNET-seq)⁷¹, this method became possible to use in metazoans. Since NET-seq is still based on immunoprecipitation, the data from it can be influenced by antibody quality and crosslinking time.

2) Methods measuring Pol II enzymatic activity

Global run-on sequencing (GRO-seq)⁷² was a major breakthrough in the attempt to measure Pol II speed genome-wide⁷³. Nascent RNA is extended with nuclear run-on assays while initiation is prohibited. Transcription can either be newly initiated by physiological inducers⁷⁴ or paused with DRB and initiated by DRB removal⁷⁵. Elongation rate can be calculated by sampling at a time point after transcription induction, finding the “leading edge” of newly initiated Pol II and measuring the distance the polymerase has traveled during that time (speed = distance/time).

Finding the “leading edge” for the elongating polymerase is not trivial, as the read signal is noisy. Saponaro et al., who performed DRB/GRO-seq in human cells, identified the wave front by identifying regions with a normalized read depth of at least 3 base pairs from the transcription start site to 120 kb downstream⁷⁵. Stringent filtering was performed for quality control. To calculate elongation rates, the difference in wave front positions at two time points was divided by the number of minutes between them. Danko et al⁷⁴. (GRO-seq in human cells, following transcriptional activation by estrogen signaling or cytokine treatment) used a more sophisticated statistical inference procedure based on a three-state hidden Markov model. This model divides each gene into three regions: the upstream region prior to the transcription start site, the region corresponding to the Pol II wave, and the downstream region beyond the wave. To increase the reliability of their measurements, difference maps were utilized, which allow for the identification of true signals above background noise in a quantitative manner. To control for both technical errors and biological variability, the wave end was fit separately in multiple independent biological replicates at each time point and the elongation rate was calculated through linear regression analysis.

It should be noted that GRO-Seq relies on *in vitro* run-on, which can potentially affect the accuracy of the results. Specifically, the addition of sarkosyl to the run on buffer in order to prevent the

initiation of novel transcription events, may lead to the release of paused Pol II and the disengagement of regulatory factors bound to the polymerase. Furthermore, it is limited in the number of genes it can profile at the same time.

Another genome-wide approach is to label nascent RNA with bromouridine⁷⁶ or 4-thiouridine⁷⁷⁻⁸⁰ and follow the transcriptional wave after Pol II release with DRB/avlocipid. The modified nucleotides do not naturally occur in mRNAs. To label nascent transcripts with 4sU, cells are fed the compound and then purified via streptavidin-affinity purification after reacting with the thiol group of incorporated 4-thiouridines with 2-pyridylthio- or methylthiosulfonate-activated biotin (HPDP-biotin or MTS-biotin, respectively). Bru labeling works in a similar way. After the incorporation of the modified nucleotides into newly synthesized transcripts, labeled from unlabeled mRNAs can be distinguished and the wave front of transcription can be tracked with similar methods to GRO-seq. 4sU/BruDRB-seq allows the assessment of the steady state elongation rate of thousands of genes simultaneously. However, it is worth noting that the procedure can be highly complex, and it involves multiple steps that may introduce errors.

One of the published computational methods for estimating the elongation rate is the pipeline TERate⁸¹. To calculate speed, TERseq divides the transcription elongation distance by the transcription block release time for each selected gene. To find the wave front, it splits each intron into bins and calculates the expressed signal of each bin. It then randomly selected bins within the TSS proximal region as "expressed bins" and bins within the TES proximal region of very long genes as "background bins." If the tag density of the expressed bin was greater than that of the background bin, it is defined as transcribed. The position of the last two continuous transcribed bins is then defined as the Pol II transcription edge, and the transcription elongation distance is calculated from the TSS to this edge.

The average Pol II speed estimation varies substantially depending on the method and system used (1.25 to 3.5 kb/min). Speed was also shown to be very variable, not only between different genes (0.37 to 3.57 kb/min) but also in the same genes in different cell lines and within a gene (accelerating 4-8 times downstream compared to upstream). This indicates that the regulation of transcriptional speed by RNA polymerase II is a highly dynamic process that occurs both between and within genes.

Total RNA-seq data has also been used to calculate Pol II speed^{82,83}. Instead of performing selection by polyA tailing, like standard RNA-seq, selective depletion of rRNAs from the total RNA of a tissue is performed. Unlike polyA sequencing, total RNA-seq retains transcripts that are still not fully transcribed, so the final pool of reads contains both mature polyadenylated RNAs and nascent transcripts.

In total RNA-seq datasets, the read coverage in the introns displays a monotonically decreasing slope from the 5' to the 3' end of the intron, with substantially higher RNA levels present in the exons. The reason these slopes exist in the introns and not the exons is a result of nascent RNA synthesis being concurrent with intron removal⁸². The 3' ends of the introns are excised and degraded shortly after the 3' splice sites are fully transcribed. This creates a characteristic saw-tooth pattern of read coverage in the actively-transcribed genes. The same pattern can be seen in nascent or "factory" RNA-SEQ⁸⁴, which captures nascently-transcribed molecules.

The rate of transcription elongation by Pol II is a major factor influencing this pattern. When RNA Pol II moves quickly through an intron, it reaches the downstream exon more swiftly, enabling co-transcriptional splicing to occur. This results in a decrease in RNA accumulation at the 5' end of the intron, which is reflected in the slope of intronic read coverage. Faster elongation rates result in shallower decreasing gradients in introns, while slower rates result in steeper gradients. The average speed of Pol II in a long intron can be estimated by fitting a linear model on the read coverage and using the estimated coefficient of the model. The calculated slopes can then be compared among different genes from the same sample or the same genes in different samples. This technique is quite noisy and best applied for relative comparisons of Pol II speed among different samples, but it gives the opportunity to compare speeds in thousands of different publicly available datasets, from different tissues, organisms and conditions.

Finally, in Chapter 2, Pol II speed is calculated for the first time using a combination of DRB pausing and reinitiation of transcription with thiouridine to cytidine conversion sequencing (TUC-seq). 4sU is incorporated in the cells in the same way as 4sUDRB-seq, but instead of using thiol-specific biotinylation, T-to-C nucleotide conversions are chemically induced. These are then detected in the sequencing data after alignment as point mutation, allowing the separation of labeled and unlabeled reads and the genome-wide detection of the transcriptional wave front without the complicated affinity purification process. The biggest challenge of this approach is correctly identifying reads from newly synthesized reads, since the rate of 4sU incorporation is low and the frequency of sequencing artifacts is high. The biochemical and computational details behind our implementation of the method are explored in more detail in the methods section of Chapter 2 of the dissertation.

1.3.2 Pol II speed regulation

The structure of the chromatin is one of the major factors affecting Pol II speed. In the eukaryotic nucleus, the DNA is packaged in chromatin, the basic unit of which is nucleosomes. They consist of 147 bp of DNA wrapped around a protein core made of histones (two heterodimers of H2A-H2B and two heterodimers of H3-H4)⁸⁵. Nucleosomes are separated from each other by a sequence of non-nucleosomal DNA known as linker DNA, the length of which ranges from ~20 to 90 bp, depending on the organism, tissue or even cell⁸⁶. Because the DNA has to fit in the limited nuclear space, nucleosomes are not in a linear state. Instead, they are compacted on top of each other into a more condensed chromatin fiber⁸⁷. If the chromatin is highly condensed, it is called heterochromatin; if it is less condensed, it is called euchromatin. Euchromatin is more transcriptionally permissive since the DNA is more accessible to Pol II when it is less densely packaged. Additionally, it is known that nucleosomes by themselves hinder transcriptional elongation³³ and they can induce polymerase pausing^{69,88}. In *Drosophila*, it was shown that they are barriers to transcription in all active genes, as Pol II stalls at their entry site⁸⁹. Tellingly, histone depletion in human cells, which promotes a more open chromatin configuration, increases Pol II speed⁹⁰.

Additionally, histone modifications can also have an effect on elongation rates by changing chromatin density. When a histone amino acid is modified, its charge can change, altering the interaction between the protein and the DNA and thus the compaction of the protein. Alternatively, some modifications do not act by directly changing the interaction between the histones and the

genome, but by allowing the binding of transcription factors that can modify proteins. Histone modifications that cause denser chromatin slow down Pol II while those that cause looser chromatin accelerate it^{91,92}. More specifically, H3 and H4 acetylation⁹¹, H4K20 methylation⁷⁶, H3K79 dimethylation^{76,83} and H2B monoubiquitylation⁹³ correlate with increased intragenic Pol II speed while H3K27 trimethylation⁹² and H3K9 dimethylation⁹² correlate with decreased Pol II transcription rate. There are also some indications that methylation of the gene body DNA itself can slow down elongation rate^{94,95}, possibly through the involvement of methyl-CpG binding proteins⁹⁶.

Not only does the structure of chromatin have an influence on the speed of Pol II, but also the sequence of the transcribed DNA has a relevant effect. Pol II can pause and even backtrack because of the DNA sequence and the resulting structure of the nascent RNA. Weaker RNA-DNA hybrids transcribed from DNA sequences rich in A-T promote pausing while GC-rich templates have fewer pauses^{97,98}. It is worth noting, however, that, in vivo, GC content has been shown to have zero⁹⁹ or negative⁷⁶ correlation with Pol II speed, hinting at a more complicated role of GC sequences in transcription kinetics. DNA sequences have also been shown to govern the positioning of nucleosomes, as nucleosomes have clear sequence preferences¹⁰⁰.

The structure of a gene, specifically the presence of introns and exons, can also affect the speed at which Pol II transcribes the gene. Transcription is faster if introns are present in the transcript⁶⁵ and slower if the transcript has a higher density of exons⁷⁶. This may be due to increased nucleosome occupancy on exons^{101–103}, which as previously described can cause Pol II to pause, and the involvement of pre-mRNA splicing in slowing down Pol II over exons, as it has been shown that Pol II pauses on splice sites in yeast¹⁰⁴, fruit flies¹⁰⁵ and humans⁷¹. The same studies have also indicated higher concentrations of polymerase at spliced exons compared to skipped exons. These findings suggest the transcription of RNA polymerase II slows down intron-exon boundaries, and this slow-down may be linked to the splicing process. This slowing may also influence co-transcriptional splicing.

Finally, a few transcription factors are also known to have an effect on the elongation rate of Pol II. Ccr4-Not and TFIIIS can stimulate Pol II elongation by rescuing it from backtracking^{106,107}. Indeed, it has been directly shown that dominant-negative TFIIIS slows down transcription by half¹⁰⁸. The direct interaction of other factors with Pol II speed is more difficult to demonstrate, since they are often involved with multiple transcriptional processes. Knockdown of Spt6, which mediates the reassembly of nucleosomes¹⁰⁹, slows down elongation rate in *Drosophila* from 1.1 kb/min to 0.5 kb/min⁶⁶. The *spt5-242* mutation of Spt5 in yeast decreases elongation rates¹¹⁰, although it has been shown to have no effect on speed in MEFs (mouse embryonic fibroblasts)¹¹¹. MYC-dependent recruitment of Spt5 increases Pol II speed^{79,112} and Spt5 dephosphorylation by the PNUTS-PP1 phosphatase slows down Pol II transcription⁵⁰. RECQL5 depletion significantly increases the elongation rate, while also increasing backtracking and pausing events⁷⁵. In the absence of Sub1, a PIC component and global regulator of Pol II phosphorylation, elongation slows down in the GAL1 gene in yeast¹¹³. Inhibition of the cyclin-dependent kinases CDK12 and CDK13 greatly reduces Pol II speed¹¹⁴. SCAF8 positively affects Pol II speed¹¹⁵ and so does Paf1C in yeast¹¹⁶. The wide range of species, conditions and technical approaches used to measure the effects of transcription factors on Pol II speed makes assessment of their relative importance very complicated.

1.3.3 Processes regulated by Pol II speed

The main method of investigating the effects of Pol II speed on co-transcriptional processes is using mutations that increase or decrease elongation speed and studying their effects on an organismal and cellular level. A lot of these mutations were discovered in a *Drosophila* genetic screen because they confer resistance to α -amanitin, a mushroom toxin that inhibits transcription, and they are mostly concentrated in the RPB1 subunit of Pol II¹¹⁷. The C4 mutation (R741H) specifically has a lower elongation rate. Its human homologue, R749H, also slows down transcription¹¹⁸. H1085Y in yeast and its human homologue H1108Y slow down transcription, whereas E1103G in yeast and its human homologue E1126G speed transcription up¹¹⁹. Multiple studies have been performed using these mutants that show a very significant effect of Pol II speed on multiple cellular processes vital for healthy cellular function. For instance, it has been established in mice that mutants with reduced elongation rate exhibit early embryonic lethality, which could indicate that proper control of Pol II speed is essential for the correct expression of developmental genes¹¹⁴. Some of the more important processes regulated by Pol II speed are analyzed below.

Splicing

Splicing was first discovered in the late seventies in two studies that showed that eukaryotic genes were split into exons and introns^{120,121}. The introns are removed or 'spliced out' of the final transcript and the exons are combined together. This explains the confusing previous finding that nuclear mRNAs were much longer than cytoplasmic mRNAs, despite having the same beginning and end sequences. There are multiple sequences that define an intron, like the donor site at the 5' exon-intron border, the branch point close to the 3' end and the acceptor site at the 3' intron-exon border. The donor and acceptor sites are known as **splice junctions**. The splicing reaction takes place in two steps that cut the intron in the splice junctions and join the exons in two sequential transesterification reactions¹²². The process is conducted by the spliceosome, a large ribonucleoprotein complex¹²³.

In constitutive splicing, the pre-mRNA is always spliced in the same way. On the other hand, alternative splicing is a process in which a single mRNA is spliced in different ways. This increases transcriptomic and proteomic diversity because a higher number of proteins is synthesized compared to the number of genes¹²⁴. There are many variants of alternative splicing, including exon skipping, intron retention and alternative donor/acceptor site. It is estimated that ~95% of human genes undergo alternative splicing, in a tissue and stimulus-specific manner¹²⁵.

Several studies that use mutants with different elongation speeds have indicated that Pol II speed and constitutive splicing efficiency are inversely correlated in budding yeast^{9,126} and *Drosophila*¹²⁷. Similar to constitutive splicing, changes in transcription speed seem to also have an effect on alternative splicing events^{119,128-132}.

Transcriptional initiation and termination

Beyond alternative splicing, transcriptome variability is also increased by the alternative choice of transcription start sites (TSSs)¹³³ and transcription end sites (TESs)¹³⁴. Transcript isoforms with different TSSs and TESs can also vary in their localization, stability and translational efficiency¹³⁵. Pol II speed affects TSS preference in budding yeast promoters. Increased elongation speed shifts

initiation upstream, whereas decrease in speed shifts initiation downstream^{126,136}. This means that changes in Pol II speed could possibly have downstream effects on transcriptome composition or translation activity¹³⁷. Pol II speed has also been shown to affect both the duration¹³² and the location¹³⁸ of polymerase pausing.

Additionally, differences in transcriptional termination have been demonstrated in Pol II speed mutants, as would be expected from the “torpedo” model. Slowing down transcriptional elongation shifts termination upstream and, conversely, accelerating elongation reduces the efficiency of termination, shifting it downstream^{132,139}. Therefore, elongation acceleration can increase transcriptional read-through, potentially affecting the gene expression of neighboring genes and increasing transcriptional noise.

Regulation of chromatin structure and gene expression

As discussed above, histone modifications can have an impact on elongation speed. There are several studies that show the reverse is also true: Pol II speed affects the pattern of histone modifications on the parts of the chromatin that is getting transcribed. H3K4 methylation patterns change in response to elongation speed¹⁴⁰ and the monoubiquitylation of H2B is significantly correlated with the speed of Pol II⁹³. Both of these modifications play important roles in transcriptional regulation and therefore in the regulation of gene expression.

The effect of transcription speed on RNA expression levels is unclear. There has been some evidence from NET-SEQ⁹⁹ and GRO-SEQ⁷⁴ experiments that genes with high speed also exhibit high expression rate. Nonetheless, other studies have found no correlation between elongation rate and expression levels^{57,65}.

Transcriptional fidelity

One potential effect of changes in Pol II speed is altering transcriptional fidelity, the accuracy with which genetic information is transcribed from DNA to RNA. The speed of synthesis by RNA^{141,142} and DNA¹⁴³ polymerases seems to be inversely correlated with its accuracy. However, there is no direct genome-wide experimental measurement of Pol II speed and error rate at the same time, as simultaneous measurement of both of these factors is difficult.

1.4 Transcription elongation errors

The fidelity of transcription is important since errors in transcription can lead to non-functional RNAs or to mRNAs that get translated to proteins that are truncated or dysfunctional¹⁴⁴. As important as transcriptional accuracy is, Pol II still makes mistakes, with an estimated error rate of less than 1 in 10000 in prokaryotes and eukaryotes^{145,146}. The first part of this section is a review of the known cellular mechanisms that correct errors in transcription. The second part is an exploration of the known effects caused by mistakes that evade correction and their potential consequences on health and aging. Finally, the third part is a description of the methods that have been used both historically and recently to estimate transcription error rate.

1.4.1 Transcription errors and correction mechanisms

Transcriptional fidelity is a result of three processes: correct substrate selection, proofreading and non-efficient extension of transcripts with misincorporated nucleotides. After the RNA translocates from the substrate site to the product site, the nucleotide selection takes place in two steps. First, the nucleotide binds to the open active center of Pol II. If it is complementary to the DNA base, it is delivered to the insertion site¹⁴⁷ of the enzyme. Catalysis of a phosphodiester bond is performed and the active site closes through the folding of a specific structural domain of the polymerase called the trigger loop¹⁴⁸. This domain is very important for fidelity. When the correct nucleotide binds to the active center, the trigger loop folds and catalyzes the nucleotide's incorporation. If the nucleotide is not complementary to the DNA or if it is a complementary deoxynucleotide, then it cannot induce correct folding of the trigger loop¹⁴⁹.

If misincorporation occurs, the lack of proper alignment between the ribonucleotide and the DNA template causes the fraying of the nucleotide from the template¹⁵⁰. Pol II pauses and backtracks³¹ by at least 1 bp. Transcription stops until the mismatch is cleaved, allowing a second transcription of the previous wrongly transcribed base. There are two cleavage mechanisms: intrinsic cleavage, in which the active site of the polymerase itself cleaves the mismatched nucleotide¹⁵¹⁻¹⁵³, and factor-assisted cleavage, involving proteins that stimulate cleavage like GreA and GreB in bacteria^{154,155}, and TFIIIS in eukaryotes¹⁵⁶.

Beyond misincorporations, Pol II also commits frameshifts (either insertions or deletions). Frameshifts are more common in homopolymeric A/T regions¹⁵⁷, but they are rarer than misincorporations, since they tend to be more destabilizing to the template-RNA hybrid. They can potentially disrupt cellular homeostasis in a more significant way, since mRNAs with insertions or deletions commonly contain premature termination codons (PTCs) and they are either eliminated by the NMD pathway or translated as truncated proteins.

1.4.2 Physiological consequences of transcription elongation errors

If, despite these quality control mechanisms, errors escape correction, they can have a profound effect on cellular phenotype. After all, transcription errors are amplified with translation, as one RNA molecule can produce multiple proteins¹⁵⁸. Additionally, even though errors happen very rarely and have an impact even more rarely, transcription takes place constantly on gigantic scales. This

means that even a very infrequent event can occur a large number of times in the lifespan of an organism.

Since misincorporations can change the amino acids in important catalytic or binding domains, they can alter or completely deactivate normal protein function. This is especially important for proteins with a long half-life, as the consequences of the mistake can persist long after it originally occurred. Even a single error can have dire consequences, as shown by a study in which O6-methylguanine-induced misincorporation of a single uridine in the *TP53* gene caused impaired apoptosis and cell cycle arrest in almost 15% of the cells¹⁵⁹. The odds of one error having long-term physiological consequences are even further increased in long-lived, non-replicating cells such as neurons.

Significant issues can also be caused by the cumulative effects of less important transcriptional errors, as has now been shown in multiple studies. Yeast mutants with increased transcription error rate suffer from proteotoxic stress and reduce cellular lifespan⁵. Errors can also result in splicing defects¹⁶⁰, mutagenesis¹⁴⁴ and tumorigenesis¹³. Misincorporation-caused Pol II pausing can also result in physical blocking of transcription of important genes and conflict with other cellular mechanisms¹⁶¹. Finally, RNA errors are a contributor to molecular noise, impacting the homeostasis of the proteome and thus the homogeneity of a tissue¹⁶².

Transcriptional infidelity has also been connected to neurodegenerative aging-related diseases, since transcription errors in specific genes expressed in the brain can cause the production of disease-related protein aggregates and apply pressure to the mechanisms that control protein quality^{11,12}. In addition, it has been shown that aging increases the transcriptional error rate in yeast cells⁵ and these errors can influence the aggregation and degradation of proteins which contribute to aging-associated diseases in humans. Given that disruption of proteostasis is one of the recognized hallmarks of aging⁴, the error rate of transcription is of special interest in regard to aging.

1.4.3 Measurements of transcription errors

The accurate measurement of the error rate of transcription and its changes in various conditions, including aging, is important. It would help with the determination of the effects this has on the proteome and on the subsequent phenotype. This could allow the identification of specific genes that are disproportionately affected and thus potential therapeutic targets. Additionally, an accurate error rate estimation could provide insight into the factors, genetic or environmental, that contribute to transcriptional infidelity and allow their investigation in the context of aging.

The earliest error rate measurement for RNA polymerases was estimated *in vitro* in 1975 by Springgate and Loeb based on the rate of misincorporation of radioactively-labeled nucleotides¹⁶³. They used repeating dinucleotide templates and purified bacterial polymerase for the measurement. Even though the study had significant issues (no elongation factors or additional regulatory proteins, repeating template which increases transcription errors), it was a milestone for many years, since measuring error rates *in vivo* was an intractable problem for existing technology at that time.

Before the advent of NGS techniques, *in vivo* measurements were limited to single-cell organisms. The first ones were performed using the *lac* operon¹⁶⁴. The *lacZ* gene was modified so that it

contained a premature stop codon, hence the only way for LacZ to be functional was if a transcriptional mistake would change the premature stop codon to a functional amino acid. The activity of beta-galactosidase was then measured after induction with LacI. The authors argued that the effect of translational mistakes would be negligible because of the extreme polarity of the mutation. Additionally, multiple mistakes in translation would have to occur in the same cell in the same codon to generate a functional beta-galactosidase tetramer. The RNA polymerase error rate indicated by the assay was 1.4×10^{-4} . The assay required both a lot of assumptions about translation and transcription that are not completely accurate as well as very high precision in the measurement of both LacZ activity and number of cells. However, despite these limitations, it is still used to this day to compare error rates in different bacterial mutants^{165,166} and calculate error rate in yeast, wild-type¹⁶⁷ or mutant^{168,169}.

Another non NGS-based approach used a Cre/Lox *gal* transcription fidelity reporter system in *E. coli*¹⁷⁰. In this system, the Cre recombinase is rendered inactive due to a missense mutation in the active site of Cre. In the event of a transcription misincorporation error in the mutation site, Cre is restored and translated into a functional tetramer. Its activity can be measured, since the functional Cre converts a *gal*- mutant gene to *gal*+, allowing the growth of Gal positive colonies on selected growth media. The same approach in yeast provided valuable information about fidelity mutants in eukaryotic Pol II¹⁷¹.

More recently, there have been attempts to measure transcriptional error rate with RNA-seq¹⁷². This is still technically very challenging, as the reverse transcription process necessary for converting RNA to cDNA and the sequencing itself are error-prone and thus it is very difficult to distinguish transcriptional errors from sequencing artifacts^{173,174}. One way around this problem involved ligating a randomized barcode on each RNA fragment, followed by multiple rounds of cDNA sequencing. This allows for the creation of a consensus read among sequences with the same barcode, allowing distinction of artifacts from true errors. The authors termed this method Rep-seq¹⁷⁵. Its main disadvantage, the very low efficiency, was resolved with the development of CirSeq, which generates tandem cDNA repeats from a single RNA fragment with a rolling circle polymerase¹⁷⁶⁻¹⁷⁸. CirSeq was recently improved by changing the RNA fragmentation strategy, which was artificially increasing the detected error rate¹⁶². Another approach based on CirSeq is called ARC-seq (Accurate RNA Consensus sequencing), in which a barcode is added to each RNA molecule and multiple cDNA copies of each molecule are generated¹⁷⁹. While tandem repeat methods are a massive improvement in determining error rate compared to what was previously available, they still suffer from specific limitations. First, there is no way to distinguish between RNA editing changes and transcription errors. Second, rare *de novo* mutations are impossible to correct for. Finally, and most importantly, CirSeq is a protocol that requires specialized expertise and experience and has not been widely applied, so there are limited available datasets.

Finally, NET-Seq has also been used for error rate estimation. Pausing has been linked to misincorporation, by comparing the error rates between the most recently transcribed nucleotide and the rest of the RNA-DNA hybrid¹⁶¹. As opposed to CirSeq, NET-Seq data is routinely generated for other applications and there are a lot of publicly available datasets. The sequencing is still error-prone, making this method more suitable for relative instead of absolute comparisons. NET-Seq can be further optimized by adding an extra RNase footprinting step to distinguish pausing

from backtracking (RNET-Seq)¹⁸⁰, but its sensitivity relies on sequencing small RNAs which are difficult to map to the genome. This limitation restricts the use of RNET-Seq.

In Chapter 3, we will introduce a novel computational method, `scErrorRate`, for the quantification of transcription elongation errors through analysis of single-cell RNA-seq data. This approach has the potential to be applied to a wide range of datasets from various individuals, cell types and health and disease states. It has the potential to be applied to thousands of existing datasets, which could lead to significant advances in the field of transcriptional fidelity.

1.5 Single-cell RNA sequencing and unique molecular identifiers

In order to fully understand and effectively utilize scErrorRate, it is important to have a thorough understanding of single-cell RNA sequencing and unique molecular identifiers. By explaining scRNAseq and UMIs in this chapter, we can provide the necessary background knowledge for the method and its potential applications.

1.5.1 Overview of commonly used scRNAseq methods

The field of transcriptomics has greatly expanded our knowledge of RNA abundance in various species, tissues, ages and conditions. For most of its existence, the field was dominated by bulk techniques (microarrays and RNA-seq), which involve measurements of RNA abundance from thousands of cells. However, in the last few years, several methods have been developed to profile the transcriptome at the level of a single cell^{181–185}. Single cell RNA sequencing (single-cell RNA SEQ/scRNAseq) has allowed investigations that were until recently impossible, providing insight into the transcriptional heterogeneity among the cells of the same tissue and the changes in transcription during cellular differentiation, as well as allowing the sorting of cell types *in silico*. By now, multiple other single cell technologies have been developed (measuring the epigenome, the proteome and the metabolome with a single cell resolution), but scRNAseq remains the most popular technique, with a wide variety of applications and approaches.

There are several methods to calculate RNA abundance on a single-cell level, but they all share common features. Every approach begins with a cell suspension and produces a count matrix. In between, they share the following steps:

1. Cells are isolated from a tissue or cell culture and lysed. Cell isolation is usually well-based¹⁸⁶, plate-based^{183,187,188} or droplet-based¹⁸⁴.
2. RNA is captured, usually with either polyA selection or rRNA depletion to enrich the molecules of interest. After capture, mRNA is reverse-transcribed to cDNA.
3. The cDNA is fragmented, amplified and tagged with sequencing adapters. It is now ready for sequencing.
4. The reads in the raw files produced from sequencing are aligned or pseudoaligned. The transcript count for all cells is then quantified, producing the count matrix.

One of the main differences between scRNAseq methods involves whether they include Unique Molecular Identifiers (UMIs) or not. UMIs are short, random sequence labels¹⁸⁹. A large combination of them can be ligated to the cDNA produced after reverse transcription, leading to the generation of a library where every molecule has a distinct nucleotide sequence. Thus, PCR copies generated from the same molecule can be tracked (duplicates) and the amplification bias caused by PCR can be removed. The data is analyzed using bioinformatics tools to identify and count the number of times each UMI is present in each cell and subsequently correct for it. UMI-based protocols are more popular since they help reduce background noise and improve data accuracy. However, most of them do not provide full transcript coverage, which is necessary for certain studies (for example,

alternative splicing analysis), since they require the incorporation of the barcode sequence through the reverse transcription primer. As a consequence, their detection is limited to the end of the sequence where they were placed, 3' or 5'. Some examples of non-UMI protocols are Smart-Seq2 and CEL-Seq¹⁹⁰, while some examples of UMI protocols are Quartz-Seq2¹⁹¹, Drop-Seq¹⁸⁴ and Chromium (10x genomics)¹⁸⁵. A recently published method called SMART-Seq3¹⁸⁷ combines UMI counts with full transcript coverage, allowing both isoform-level analysis and the mitigation of the side effects of PCR amplification.

1.5.2 Unique molecular identifiers

All single-cell data analyzed in this thesis were generated through 10x sequencing, a UMI-based protocol. Since the methodological approach we developed to estimate transcription errors requires UMIs, a deeper insight into UMIs is needed to allow a better understanding of the details of our analysis.

UMIs are short nucleotide sequences used to uniquely mark every molecule in a library, providing an increase in sequencing accuracy. UMIs are used in various sequencing methods, especially those in which correction of PCR duplicates is necessary. If PCR duplicates are not accounted for, they can falsely increase coverage, creating significant issues in many applications. Not using PCR at all is not an option for most sequencing applications, since by duplicating each DNA fragment multiple times it increases the available pool of molecules, providing sufficient coverage of the transcriptome. If it is preceded by UMI ligation, then identification of duplicates becomes a simple affair. Each UMI corresponds to one molecule, so all PCR duplicates have the exact same barcode, facilitating correction through computational methods.

UMIs have been used in many applications. Even though tagging selected genes with random barcodes had been done before^{192,193}, UMIs were first used genome-wide for DNA-Seq in 2012 for digital human karyotyping¹⁸⁹. A few years later, a different team developed a UMI-based method for *de novo* detection of mutations in plasma cell-free DNA¹⁹⁴. UMIs were subsequently used to distinguish rare sequence variants from sequencing artifacts in virology and genomics^{195,196}. Bulk RNA-seq applications of UMIs include the removal of PCR duplicates¹⁹⁷ and the identification of undiscovered sequencing artifacts¹⁹⁸.

As with bulk RNA-seq, the scRNAseq read counts that are produced after sequencing and alignment may contain PCR duplicates. This issue is more impactful in single-cell sequencing because of the smaller initial library size. These duplicates can hide differences in gene expression between cells, generate differences where none exist and negatively impact the accuracy of the data. The use of UMIs allows for the identification and removal of PCR duplicates, resulting in more reliable expression data. This is crucial for the correct interpretation of the results.

1.6 Aims of the Project

As analyzed so far, the control of polymerase II elongation rate and the fidelity of transcription are critical for proper cellular function. However, the impact of these mechanisms on aging and longevity remains largely unexplored. In this dissertation, we aim to investigate the effects of aging-associated changes in transcriptional elongation and transcriptional fidelity on aging and longevity.

Firstly, we examined how the average elongation rate is affected by aging. To do this, we used a combination of sequencing experiments to measure the rate of Pol II elongation in cells from young and old individuals. Our results show that the average elongation rate increases globally with age in various organisms and tissues, indicating that aging may dysregulate the proper function of transcription.

Next, we examined the effects of changes in Pol II speed on downstream co-transcriptional processes. Using a variety of molecular and computational techniques, we found that changes in Pol II elongation rate can impact the stability and splicing of RNA transcripts, as well as the efficiency of protein synthesis. These findings suggest that alterations in Pol II elongation rate may contribute to the decline in protein synthesis and cellular function that occurs during aging.

Finally, we developed a novel method, *scErrorRate*, to compare changes in transcriptional fidelity using single cell RNA-seq data. This method allows us to measure changes in transcriptional accuracy at the single cell level, providing valuable insights into the potential causes of aging-associated changes in transcriptional elongation.

In conclusion, our study provides new insights into the role of Pol II elongation rate and transcriptional fidelity in aging and longevity. Further research is needed to fully understand the underlying mechanisms and potential interventions to delay or prevent age-related declines in these processes.

Chapter 2. The effects of aging-associated changes in transcriptional elongation on metazoan longevity

2.1 Introduction

Aging impairs a wide range of cellular processes, many of which affect the quality and concentration of proteins. Among these, transcription is particularly important, because it is a main regulator of protein levels^{199–201}. Transcriptional elongation is critical for proper mRNA synthesis, due to the co-transcriptional nature of pre-mRNA processing steps such as splicing, editing, and 3' end formation^{7,8}. Indeed, dysregulation of transcriptional elongation results in the formation of erroneous transcripts and can lead to a number of diseases^{202,203}. During aging, animal transcriptomes undergo extensive remodeling, with large-scale changes in the expression of transcripts involved in signaling, DNA damage responses, protein homeostasis, immune responses, and stem cell plasticity⁴. Furthermore, some studies uncovered an age-related increase in variability and errors in gene expression^{5,6,204}. Such prior work has provided insights into how the transcriptome adapts to, and is affected by, aging-associated stress. However, it is not known if, or to what extent, the transcription process itself affects or is affected by aging.

In this study, we used high-throughput transcriptome profiling to investigate how the kinetics of transcription are affected by aging, how such changes affect mRNA biosynthesis, and to elucidate the role of these changes in age-related loss of function at the organismal level. We document an increase in Pol-II elongation speed with age across five metazoan species, a speed reduction under lifespan-extending conditions, and a causal contribution of Pol-II elongation speed to lifespan. We thus reveal an association of fine-tuning Pol-II speed with genome-wide changes in transcript structure and chromatin organization.

2.2 Results

The translocation speed of elongating RNA polymerase II (Pol-II) can be measured using RNA sequencing (RNA-seq) coverage in introns. This is because Pol-II speed and co-transcriptional splicing are reflected in the characteristic saw-tooth pattern of read coverage, observable in total RNA-seq or nascent RNA-seq measurements^{56,82}. Read coverage generally decreases 5' to 3' along an intron, and the magnitude of this decrease depends on Pol-II speed: the faster the elongation, the shallower the slope^{57,77,83}. High Pol-II speeds result in fewer nascent transcripts interrupted within introns at the moment when the cells are lysed. Thus, by quantifying the gradient of read coverage along an intron, it is possible to determine Pol-II elongation speeds at individual introns (Figure 2.1a,b). Note that this measure is only weakly associated with the expression level of the transcript (Supplementary Table 2.1).

To monitor how the kinetics of transcription changes during aging, we quantified the distribution of intronic reads resulting from RNA-seq in five animal species: the worm *C. elegans*, the fruit fly *D. melanogaster*, the mouse *M. musculus*, the rat *R. norvegicus*, and humans *H. sapiens*, at different adult ages (Supplementary Table 2.2 and Materials and methods), and using diverse mammalian tissues (brain, liver, kidney, whole blood), fly brains, and whole worms. Human samples originated from whole blood (healthy donors, age 21-70), and from two primary human cell lines (IMR90, HUVEC) driven into replicative senescence.

After filtering, we obtained between 518 and 7994 introns that passed quality criteria for reliable Pol-II speed quantification (Materials and Methods). These different numbers of usable introns mostly result from inter-species variation in intron size and number, and to some extent from variation in sequencing depth. To rate the robustness of Pol-II speed changes across biological replicates, we clustered samples based on their 'speed signatures', i.e. on the detected elongation speeds across all introns that could be commonly quantified across each set of experiments. We observed largely consistent co-clustering of samples from the same age across species, whereas young and old samples mostly separated from each other (Supplementary Figure 2.1). This suggests that age-related speed changes were consistent across biological replicates and reliably quantifiable in independent measurements.

We observed an increase of average Pol-II elongation speed with age in all five species and all tissue types examined (Fig. 2.1 and Supplementary Fig. 2.2). Changes in Pol-II speed did not correlate with either the length of the intron or with its position within the gene (Supplementary Table 1). The observed increase in Pol-II elongation speed was even more pronounced after selecting introns with consistent speed changes across all replicates (i.e., always up or down with age; Supplementary Fig. 2.3). This result is non-trivial, because our analysis also revealed introns with a consistent reduction in Pol-II speed.

In order to confirm our findings with an orthogonal assay, we monitored transcription kinetics in IMR90 cells using 4sU-labeling of nascent RNA. After inhibiting transcription with 5,6-dichloro-1- β -D-ribofuranosyl benzimidazole (DRB), we conducted a pulse-chase-like experiment quantifying 4sU-labeled transcripts at four time points after transcription release (i.e., at 0, 15, 30 and 45 min). This enabled us to quantify Pol-II progression into gene bodies (see Methods for

details) and confirmed our results based on intronic slopes using proliferating (young) and senescent (old) IMR90. Pol-II speed measurements from the 4sU-based assay showed significant correlation with those from the slope-based assay (Fig. 2.1d), with Pol-II speed increasing on average in both approaches (Fig. 2.1e, Supplementary Fig. 2.4). Note that, although many individual genes showed a decrease in elongation speed with aging in both assays, the majority exhibited increased speed.

To assess whether known lifespan-extending interventions, inhibition of insulin/insulin-like growth factor signaling (IIS) and dietary restriction, affected Pol-II speeds, we sequenced RNA from IIS mutants, using *daf-2* mutant worms at day 14 and fly brains from *dilp2-3,5* mutants at day 30 and day 50, as well as hypothalamus from aged wild type and *IRS1*-null mice. We also sequenced RNA from kidney and liver of dietary restricted (DR) and ad libitum-fed mice. In all comparisons, except *IRS1*-null mice and livers from 26 months old DR mice, lifespan-extending interventions resulted in a significant reduction of Pol-II speed. Pol-II elongation speeds thus increased with age across a wide range of animal species and tissues, and this increase was, in most cases, reverted under lifespan-extending conditions (Fig. 2.1).

Although Pol-II speed changed consistently with age across replicates (Supplementary Fig. 2.1), we did not observe specific classes of genes to be affected across models. To determine whether genes with particular functions were more strongly affected by age-related Pol-II speed changes, we performed gene set enrichment analysis on the 200 genes with the highest increase in Pol-II speed during aging in worms, fly brains, mouse kidneys and livers, and rat livers. Only very generic functional classes, such as metabolic activity, were consistently enriched across three or more species (Supplementary Fig. 2.5). Thus, no specific cellular process appeared to be consistently affected across species and tissues. Next, we examined age-associated gene expression changes of transcription elongation regulators. We observed that some regulators (e.g., *PAF1*, *THOC1*) were consistently downregulated across species during aging (Supplementary Fig. 2.6), which was also confirmed using gene set enrichment analysis (Supplementary Fig. 2.7). These expression changes potentially represent a compensatory cellular response to a detrimental increase in transcriptional elongation speeds.

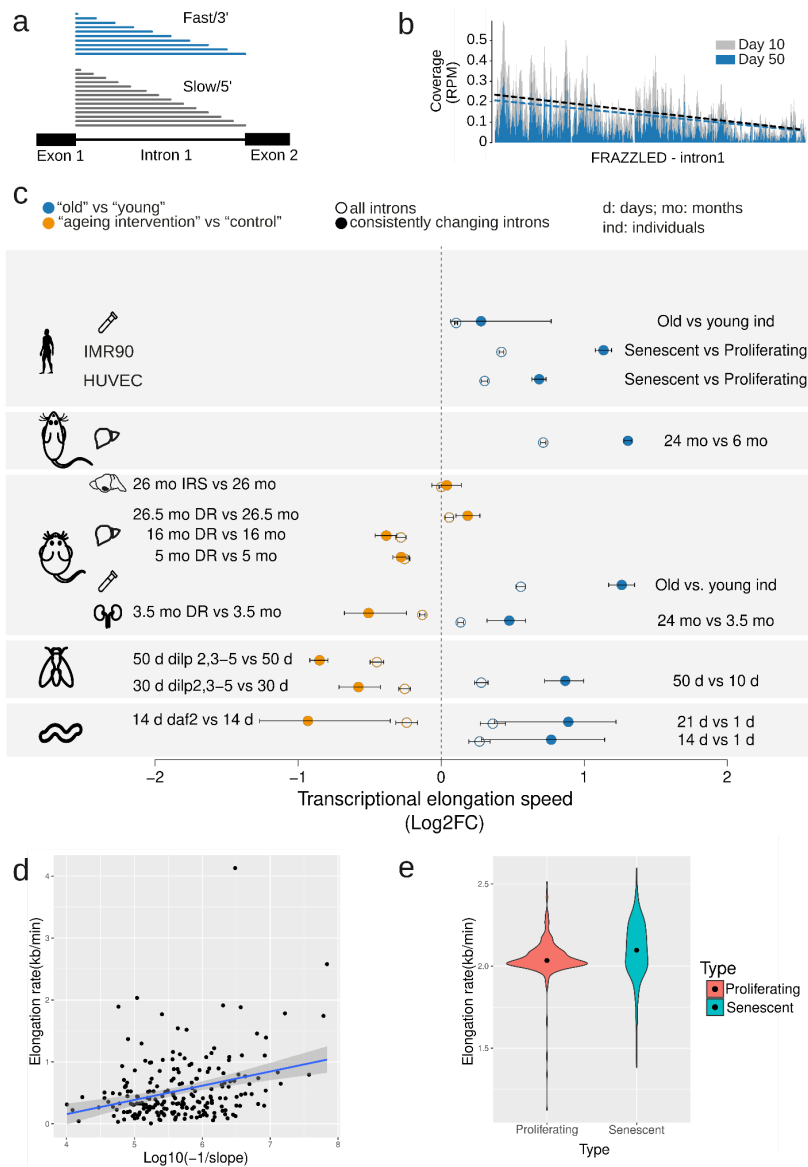


Figure 2.1: Pol-II elongation speed increases with age and is slowed-down by reduced insulin signaling and dietary restriction (DR) in multiple species. (a) Schematic representation of read coverage along introns in total RNA seq. Intronic reads represent transcriptional production at a given point in time. A shallower slope of the read distribution is a consequence of increased Pol-II elongation speed. **(b)** Exemplary read distribution in the FRAZZLED intron 1 with coverage in reads per million (RPM) for *D. melanogaster* at age day 10 (gray) and day 50 (blue). **(c)** Log2 fold change of average Pol-II elongation speeds in worm (whole body), fruit fly (brains), mouse (kidney, liver, hypothalamus, blood), rat (liver), human blood, HUVECs: umbilical vein endothelial cells; IMR90: fetal lung fibroblasts). Error bars show median variation $\pm 95\%$ confidence interval (Wilcoxon signed rank test). Empty circles indicate results using all introns passing the initial filter criteria, while full circles show results for introns with consistent effects across replicates. Number of introns considered (n) ranged from 518 to 7994. **(d)** Transcriptional elongation speed estimate from 4sUDRB-seq in IMR90 cells versus intronic slopes for 217 genes for which elongation speed could be estimated using both assays. Each dot represents one gene (Pearson correlation=0.313, p-value=2.5e-06).

(e) Distributions of elongation speeds in IMR90 cells based on 4sUDRB-seq. The black dot indicates the average speed. The difference between speeds is statistically significant (paired Wilcoxon test, p -value = $2.13e-10$). The same genes (464 genes) were used for both conditions (see Methods for details).

To determine if changes in Pol-II speed are causally involved in the aging process, we used genetically modified worm and fly strains carrying point mutations in a main Pol-II subunit that reduce its elongation speed (*C. elegans*, *ama-1* (m322) mutant²⁰⁵; *D. melanogaster*, *RpII215C4* mutant²⁰⁶). We sequenced total RNA from wild type and “slow” Pol-II mutant worms (whole animal at day 14) or fly heads, at day 10 and 50. Measurements of elongation speeds confirmed the expected reduction of average Pol-II speeds in both *C. elegans* *ama-1* (m322) and *D. melanogaster* *RpII215C4* (Fig. 2.2a). To assess whether Pol-II speed and its associated maintenance of transcriptional fidelity also affected aging of the whole organism, we measured survival of these animals. Slowing down Pol-II increased lifespan in both worms and fruit flies (median lifespan increase of ~20 % in *C. elegans* and in ~10 % *D. melanogaster*; Fig. 2.2b and Supplementary Fig. 2.8a). CRISPR/Cas9 engineered reversal of the Pol-II mutations in worms restored lifespan essentially to wild-type levels (Supplementary Fig. 2.8b). Furthermore, mutant worms displayed higher pharyngeal pumping rates at older age compared to wild type worms, suggesting that healthspan was also extended by slowing down Pol-II elongation speed (Supplementary Fig. 2.9).

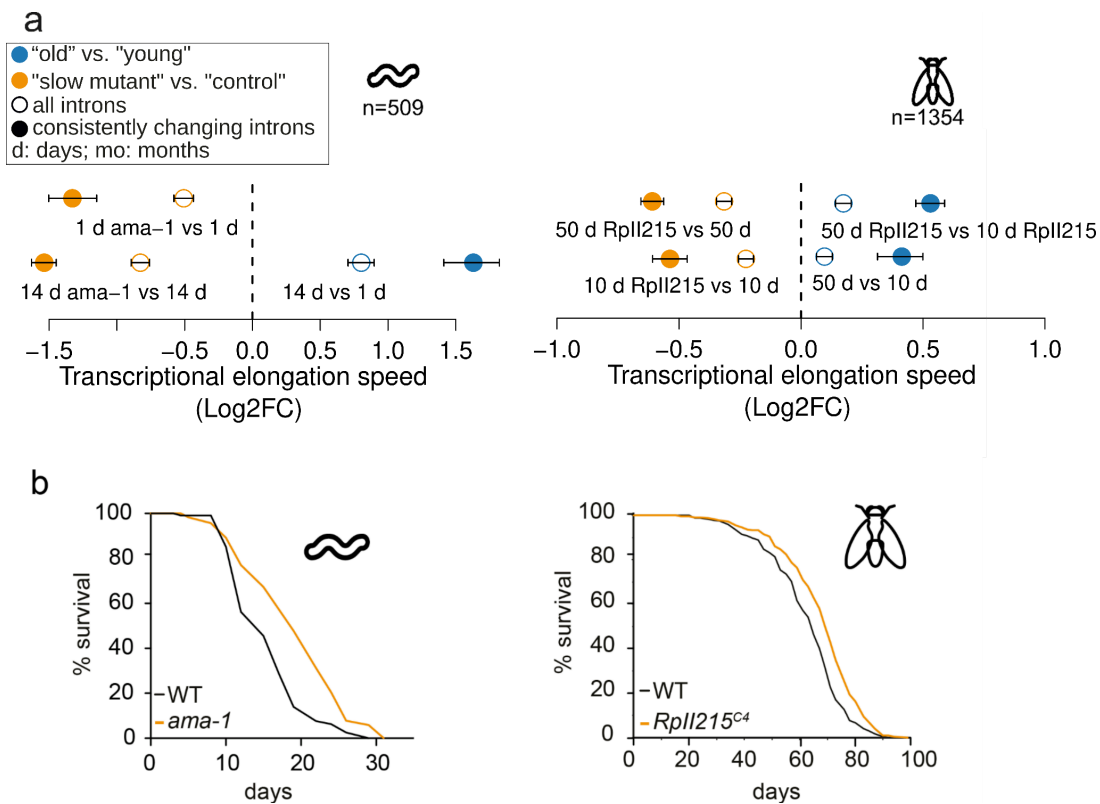


Figure 2.2: Molecular and lifespan effects of reduced Pol-II elongation speed in *C. elegans* and *D. melanogaster*. (a) Differences of average Pol-II elongation speeds between Pol-II mutant and wild-type worms (left) and flies (right), and changes of average Pol-II elongation speeds with age in flies (right). Error bars show median variation $\pm 95\%$ confidence intervals. All average changes of Pol-II elongation speeds are significantly different from zero ($P < 0.001$; paired Wilcoxon rank test). Empty circles indicate results using all introns passing the initial filter criteria, while full circles show results for introns with consistent effects across

replicates. **(b)** Survival curves of worms with *ama-1* (*m322*) mutation (**left, replicate 1**) and flies with *Rpll215^{C4}* mutation (**right, averaged survival curve**); worms 4 replicates, flies 3 replicates. Animals with slow Pol-II have a significantly increased lifespan (+20 % and +10 % median lifespan increase for *C. elegans* (n=120, P < 0.001, log rank test) and *D. melanogaster* (n=220, P < 0.001, log rank test), respectively).

Optimal elongation rates are required for fidelity of alternative splicing^{119,207}: for some exons, slow elongation favors weak splice sites that lead to exon inclusion, while these exons are skipped if elongation is faster^{7,118,129}. Faster elongation rates can also promote intron retention leading to the degradation of transcripts via nonsense-mediated decay (NMD)²⁰⁸ and possibly contributing to disease phenotypes²⁰⁹. Therefore, we next quantified changes in splicing. The first measure we used was splicing efficiency, which is the fraction of spliced reads from all reads aligning to a given splice site⁹. In most datasets, from total and nascent RNA-seq, we observed an increase of the spliced exon junctions relative to unspliced junctions during aging, and a decrease of the percent spliced junctions under lifespan-extending conditions (Fig. 2.3a). Consistent with earlier work¹³², we observed more spliced transcripts under conditions of increased Pol-II speed, i.e. greater splicing efficiency. For co-transcriptional splicing to occur, Pol-II first needs to transcribe all parts relevant to the splicing reaction (i.e., 5' donor, branch point, 3' acceptor), which are located at the opposite ends of an intron²¹⁰. Our data suggest that accelerated transcription shortens the interval in which splicing choices are made, thus shortening the time between nascent RNA synthesis and intron removal.

Accelerated transcription and splicing carries the risk of increasing the frequency of erroneous splicing events, which has been associated with advanced age and shortened lifespan²¹¹⁻²¹⁴. It is non-trivial to deduce whether a specific splice isoform is the product of erroneous splicing or created in response to a specific signal. Simply checking if an observed isoform is annotated in some database can be problematic for multiple reasons. For instance, most databases have been created on the basis of data from young animals or embryonic tissue. Thus, a detected isoform that only may be functionally relevant in old animals will not be reported in such databases. Moreover, an annotated isoform might be the result of erroneous splicing if its expression is normally suppressed at a particular age or cellular context. We therefore based our analysis on the notion that extremely rare isoforms (rare with respect to all other isoforms of the same gene in the same sample) are more likely erroneous than frequent ones^{215,216}. We used Leafcutter²¹⁷, which performs de novo quantification of exon-exon junctions based on split-mapped RNA-seq reads. Due to its ability to identify alternatively excised intron clusters Leafcutter is particularly suitable to study rare exon-exon junctions²¹⁸. We defined rare splicing events as exon-exon junctions supported by $\leq 0.7\%$ of the total number of reads in a given intron cluster, and the gene-specific fraction of rare clusters was computed as the number of rare exon-exon junctions divided by the total number of detected exon-exon junctions in that gene. We observed that such rare exon-exon junctions often resulted from exon skipping or from the usage of cryptic splice sites (Supplementary Fig. 2.10). The average fraction of rare splicing events increased during aging in fly and worm, and this effect was reverted under most lifespan-extending conditions (Fig. 2.3b). However, we did not observe a consistent age-associated increase of the fraction of rare splice variants across all species, which may at least in part be due to the more complex organization of splice regulation in mammalian cells.

Another potential indicator of transcriptional noise is the increased formation of circular RNAs (circRNAs)^{219,220}, i.e., of back-spliced transcripts with covalently linked 3' and 5' ends²²¹. Increased Pol-II speed has previously been associated with increased circRNA abundance⁸¹. Thus, we quantified the fraction of circRNAs as the number of back-spliced junctions normalized by the sum of back-spliced fragments and linearly spliced fragments. We observed either increased or unchanged average circRNA fractions during aging, while reducing Pol-II speed also reduced circRNA formation (Fig. 2.3c). This suggests that faster Pol-II elongation correlates with a general increase of circRNAs. Nevertheless, our data does not provide evidence that increased circRNA levels directly result from increased Pol-II speed, despite it being a consequence of the overall reduced quality in RNA production.

Increased Pol-II speeds can lead to more transcriptional errors, because the proofreading capacity of Pol-II is challenged⁵. To assess the potential impact of accelerated elongation on transcript quality beyond splicing, we measured the number of mismatches in aligned reads for each gene. For this, we normalized mismatch occurrence to individual gene expression levels and excluded mismatches that were likely due to genomic variation or other artifacts (see Methods for details). We observed that the average fraction of mismatches increased with age, but decreased under most lifespan-extending treatments (Fig. 2.3d). Consistent with prior findings⁵, slow Pol-II mutants exhibited reduced numbers of mismatches compared to wild-type control levels in 3 out of 4 comparisons.

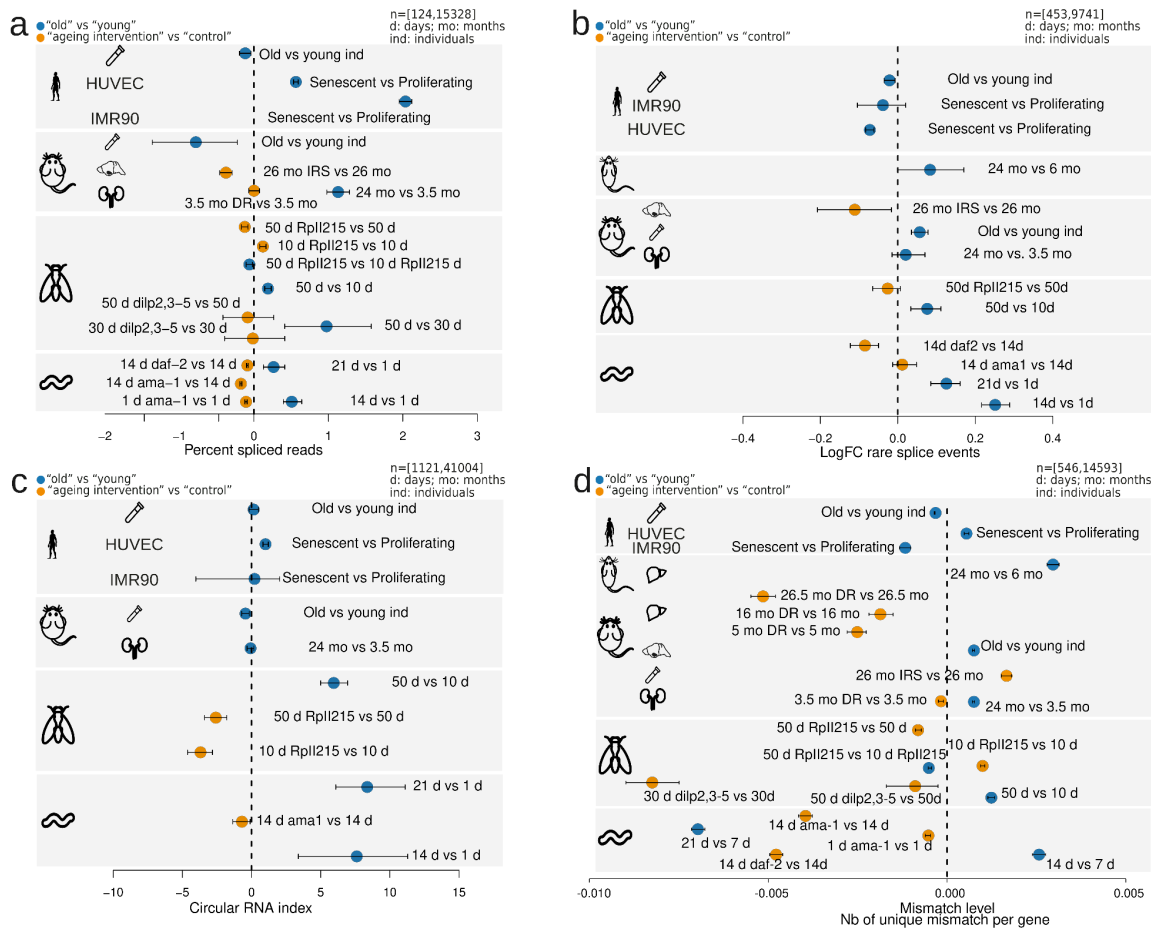


Figure 2.3: Changes in transcript structure upon aging (old vs. young; blue) and after lifespan extending interventions (orange). Error bars show median variation $\pm 95\%$ confidence interval. **(a)** Average percent changes of rare splice events (≤ 0.07 percent of total gene expression). Number of genes considered (n) ranged from 1121 to 41004. **(b)** Circular RNA index (back-spliced reads divided by sum of linear and back-spliced reads) for worms, fly heads, mouse and rat liver, human cell lines. Number of back-spliced junctions considered (n) ranged from 453 to 9741. **(c)** Average mismatch level changes. Number of genes considered (n) ranged from 546 to 14593. **(d)** Average changes of the fraction of spliced transcripts. Number of genes considered (n) ranged from 124 to 15328.

Subsequently, we explored alterations in chromatin structure as a possible cause of the age-associated changes in Pol-II speeds. Nucleosome positioning along DNA is known to affect both Pol-II elongation and splicing^{23,83,90,95}. Furthermore, aged eukaryotic cells display reduced nucleosomal density in chromatin and ‘fuzzier’ core nucleosome positioning^{17,222}. Thus, age-associated changes in chromatin structure could contribute to the changes in Pol-II speed and splicing efficiency that we observed. To test this, we performed micrococcal nuclease (MNase) digestion of chromatin from early (proliferating) and late-passage (senescent) human IMR90 cells, followed by ~ 400 million paired-end read sequencing of mononucleosomal DNA (MNase-seq). Following mapping, we examined nucleosome occupancy. In senescent cells, introns were less densely populated with nucleosomes compared to proliferating cells²²³ (Fig. 2.4a). In addition, we quantified peak ‘sharpness’, reflecting the precision of nucleosome positioning in a given MNase-seq dataset (see Methods), as well as the distances between consecutive nucleosomal summits as a measure of the spacing regularity^{223,224}. Principal Component Analysis (PCA) of the resulting signatures indicated consistent changes of nucleosome ‘sharpness’ and distances upon entry into senescence as the samples clearly separated by condition (Fig. 2.4b, c). Both measures were significantly, but moderately, altered in senescent cells (Fig. 2.4d,e): average sharpness was slightly decreased (along both exons and introns), and average inter-nucleosomal distances slightly increased in introns. In conclusion, the transition from a proliferating cell state to replicative senescence was associated with small, but significant changes in chromatin structure, involving nucleosome density and positioning-changes that were previously shown to affect Pol-II elongation^{17,23,225}.

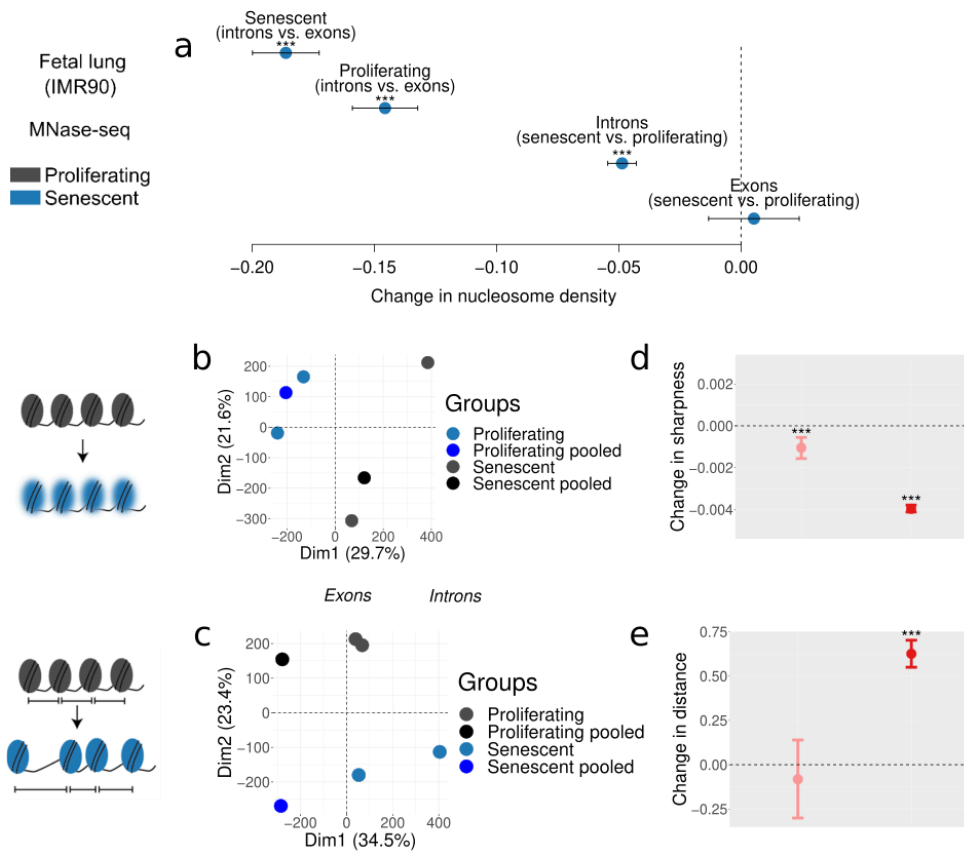


Figure 2.4: Profiling of nucleosome positioning in human cell models. (a) Average differences in nucleosome density between exons ($n=37,625$) and introns ($n=193,912$), and between proliferating and senescent cells. **(b)** Changes of nucleosome sharpness between senescent and proliferating cells in exons (left) and introns (right). **(d)** Distributions of distances between nucleosome summits between senescent and proliferating cells in exons (left) and introns (right). **(c+e)** PCA plots of nucleosome sharpness **(c)** and distances between nucleosome summits **(e)** in introns for individual samples and pooled data. All panels: Error bars show median variation $\pm 95\%$ confidence interval. Statistical significance of difference in pseudomedian distribution indicated by asterisks (***) $P < 0.001$, paired Wilcoxon rank test).

The organization of nucleosomes is severely influenced by histone availability^{90,222}. For example, histone H3 depletion reduces nucleosomal density and renders chromatin more accessible to MNase digestion²²⁶. Such global loss of histones constitutes a hallmark of aging and senescence²²⁷. Consistent with this, our senescent IMR90 and HUVECs carry significantly reduced histone H3 protein levels (Fig. 2.5a). Conversely, elevated histone levels promote lifespan extension in yeast²²², *C. elegans*²²⁸ and *D. melanogaster*²²⁹. To assess whether Pol-II elongation speed and senescence entry in human cells are causally affected by changes in nucleosomal density, we generated IMR90 populations homogeneously overexpressing GFP-tagged H3 or H4 in an inducible manner (Fig. 2.5b and Supplementary Fig. 2.11a,b). Overexpression of either histone resulted in significant reduction of Pol-II speed, confirming the causal connection between chromatin structure and transcriptional elongation (Fig. 2.5c). Pol-II speed reduction was accompanied by markedly reduced senescence-associated β -galactosidase staining in H3-/H4-overexpressing cells compared to both control (GFP-only) and uninduced cells (Fig. 2.5d). Moreover, both H3- and H4-overexpressing cells did not display p21 induction or HMGB1 depletion, both hallmarks of senescence entry, compared

to control IMR90 (Fig. 2.5e and Supplementary Fig. 2.11c). Finally, MTT assays showed that viability and proliferation were improved in H3- and, to a lesser extent, in H4-overexpressing cells compared to control cells (Fig. 2.5f). Together, these results suggest that H3/H4 overexpression decelerates Pol-II and compensates for the aging-induced core histone loss^{90,225} to restrict senescence entry.

The average speed reduction following H4 overexpression was significantly larger than that obtained upon H3 overexpression, yet H4-overexpressing near-senescent IMR90 only marginally outperformed control cells in MTT assays (Fig. 2.5f). This raises the possibility of excessive reduction in Pol-II speed negatively affecting aspects of cell function¹³¹. To address the role of nucleosome density in organismal lifespan, we used UAS-His3²²⁹ to overexpress His3, specifically in *Drosophila* glial cells using Repo-Gal4 specifically in glial cells. H3 overexpression led to significantly increased numbers of mono-nucleosomes in aged (60 day-old) compared to the wild-type fly heads (Fig. 2.5g), thus possibly compensating for age-associated loss of histone proteins. Further, H3 overexpression in glial cells increased fruit fly lifespan (Fig. 2.5h). These in vivo results are consistent with our in vitro data from IMR90, demonstrating that H3 overexpression partially reverts the aging effects on chromatin density and promotes longevity in flies. As this was linked to a reversal in Pol-II elongation speed, our findings, together with earlier ones in yeast^{222,226}, *C. elegans*²²⁸ and *D. melanogaster*²²⁹, demonstrate how the structure of the chromatin fiber likely modulates Pol-II elongation speed and lifespan.

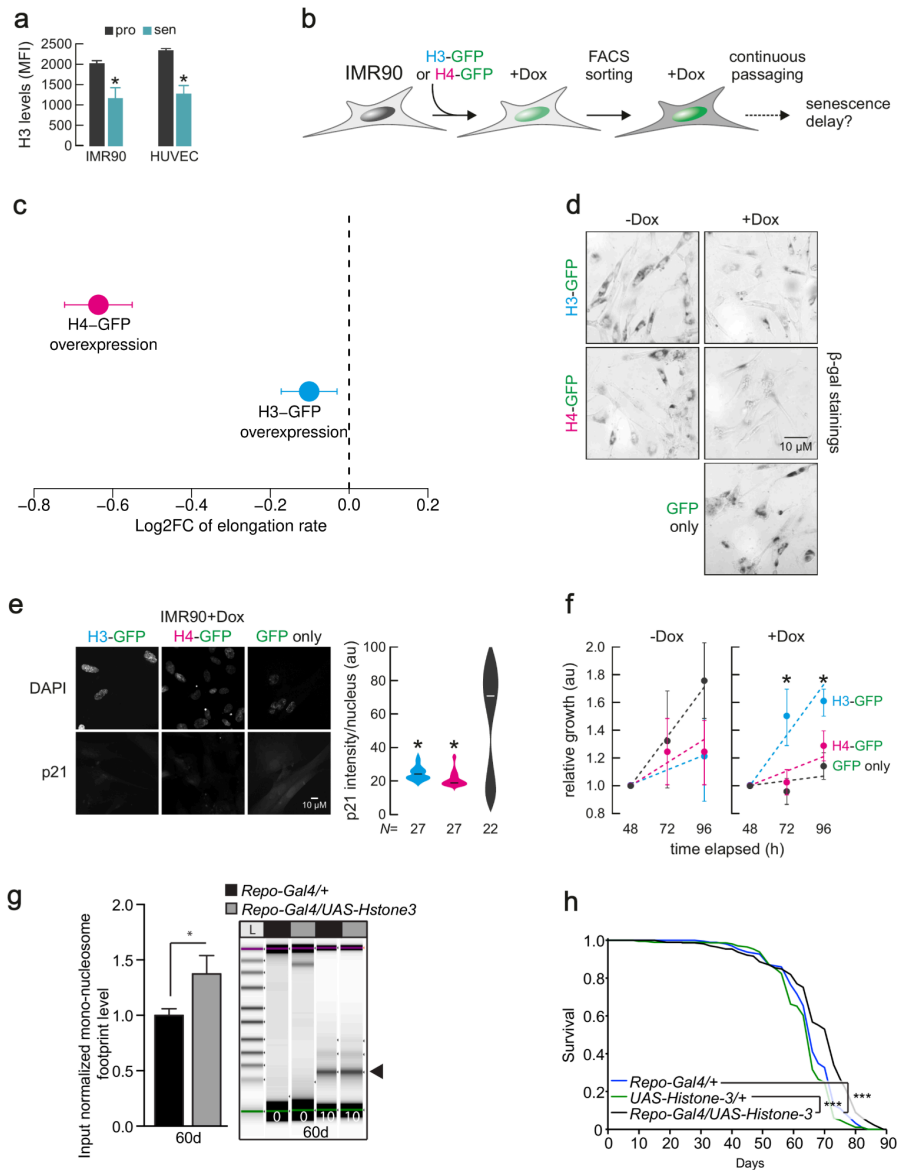


Figure 2.5 : Histone overexpression slows down entry into senescence and decreases RNA-Pol-II speed. (a) H3 protein levels in proliferating and senescent HUVEC and IMR90 cells. (b) Schematic representation of the experiment. (c) Differences of average Pol-II elongation speeds between histone overexpression mutants and wild-type IMR90 cells (derived from 1,212 introns). Error bars show median variation $\pm 95\%$ confidence interval. All average changes of Pol-II elongation speeds are significantly different from zero ($P < 0.01$; paired Wilcoxon rank test). (d) Typical images from Beta-galactosidase (β -gal) staining of H3-GFP, H4-GFP and control IMR90 cells, in the presence and absence of Doxycycline. (e) Typical immunofluorescence images of H3-GFP, H4-GFP and control IMR90 cells (left) show reduced p21 levels in histone overexpression nuclei. Violin plots (right) quantify this reduction. N specifies the number of cells analyzed per condition. (f) MTT proliferation assay. (g) Quantification of input normalized mono-nucleosome footprints (black arrowhead) between the heads of aged (60d) flies overexpressing *Histone 3* in glial cells (*Repo-Gal4/UAS-Histone 3*) and control fly heads (*Repo-Gal4/+*), significance determined by paired *t*-test ($n > 5$, * $p < 0.05$). Digests were halted after 10 min and visualized by TapeStation (Agilent) ($n > 5$). (h) Lifespan analysis of flies *Repo-Gal4/UAS-Histone3* and control flies ($+/Repo-Gal4$ and *UAS-Histone 3/+*) ($n > 100$).

2.3 Materials and methods

2.3.1 Biological materials

Eukaryotic Cell Lines. Human umbilical vein endothelial cells (HUVECs) from individual healthy donors were purchased by Lonza Inc.; human primary lung fibroblasts (IMR90) from two different isolates were obtained via the Coriell repository. All these lines were biannually checked for mycoplasma contamination and tested negative.

Human cell culture for 4sU DRB-Seq. Human fetal lung (IMR90) cells (from two donors) were grown to 80% to 90% confluence in endothelial basal medium 2-MV with supplements (EBM; Lonza) and 5% fetal bovine serum (FBS) and MEM (Sigma-Aldrich) with 20 FBS (Gibco) and 1% non-essential amino acids (Sigma-Aldrich) for HUVECs and IMR-90 respectively.

Human whole blood sample acquisition. Human samples were obtained as part of a clinical study on aging-associated molecular changes (German Clinical Trials Register: DRKS00014637) at University Hospital Cologne. The study cohort consisted of healthy subjects between 21 and 70 years of age. Whole blood samples were obtained using the PAXgene Blood RNA system (Becton Dickinson GmbH, Heidelberg, Germany) directly after informed consent.

Animal Strains Used & Animal Ethics. *Mus musculus*. Female F1 hybrid mice (C3B6F1) were generated in-house by crossing C3H/HeOJ females with C57BL/6 NCrl males (strain codes 626 and 027, respectively, Charles River Laboratories). The DR study involving live mice was performed in accordance with the recommendations and guideline of the Federation of the European Laboratory Animal Science Association (FELASA, EU directive 86/609/EEC), with all protocols approved by the Landesamt für Natur, Umwelt und Verbraucherschutz, Nordrhein-Westfalen (LANUV), Germany (reference numbers: 8.87-50.10.37.09.176, 84-02.04.2015.A437, and 84-02.04.2013.A158) and the Netherlands (IACUC in Bilthoven, NIH/NIA 1PO1 AG 17242).

Drosophila melanogaster. WT, *Rpl1215* (RRID:BDSC_3663) mutant flies and *Repo-Gal4* were obtained from the Bloomington Drosophila Stock Center (NIH P40OD018537). The *Rpl1215* allele and *Repo-Gal4* were backcrossed for 6 generations into the outbred *white Dahomey* (*wDah*) wild type background generating the *wDah*, *Rpl1215* stock, which was used for experiments. *wDah*, *dilp2-3,5* flies (RRID:BDSC_30889) and *UAS-Histone 3* (*UAS-H3*) were previously generated in the lab and backcrossed for 6 generations into the outbred *wDah* wild type background. Female flies were used for all experiments.

C. elegans strains used: AA4274 *ama-1(m322)*, *ama-1(syb2315)*, CB1370 *daf-2(e1370)*, N2 wild type.

Worm strains and demography assays. Nematodes were cultured using standard techniques at 20° C on NGM agar plates and were fed with *E.coli* strain OP50. *DR786* strain carrying the *ama-1(m322)* IV mutation in the large subunit of Pol-II (RBP1), which confers alpha-amanitin resistance, was obtained from Caenorhabditis Genetics Center (CGC)^{230,231}. *DR786* strain was then outcrossed into *N2 wild type* strain 4 times and mutation was confirmed by sequencing. 5'-3'

AGAAGGTCACACAATCGGAATC primer was used for sequencing. For each genotype, a minimum of 120 age-matched day 1 young adults were scored every other day for survival and transferred to new plates to avoid starvation and carry-over progeny. Lifespan analyses using the *C. elegans* Lifespan Machine were conducted as previously described²³². Briefly, wild-type N2 and mutant worms were synchronized by egg-prep (hypochlorite treatment) and grown on NGM-agar plates seeded with OP50 at 20°C. Upon reaching L4 stage these worms were transferred onto plates containing 0.1 g/ml 5-Fluoro-2'-deoxyuridine (FUDR) and placed into the modified flatbed scanners (35 Worms per plate). The scan interval was 30 min. Objects falsely identified as worms were censored. Time of death was automatically determined by the *C. elegans* Lifespan Machine²³². Demography experiments were repeated multiple times. For all experiments, genotypes were blinded. Statistical analyses were performed using Mantel-Cox log rank method.

Measurements of pharyngeal pumping rates in worms. Synchronized wild type and *ama-1* (m322) animals were placed on regular NGM plates seeded with OP50 bacteria on day1 and day8 adulthood and number of pharyngeal pumping rate was assessed by observing the number of pharyngeal contractions during a 10 sec interval using dissecting microscope and Leica Application Suite X imaging software. Pharyngeal pumping rate was then adjusted for the number of pharyngeal pumping per minute. Animals that displayed bursting, internal hatching and death were excluded from the experiments. Experiments were repeated three independent times in a blinded fashion, scoring a minimum of 15 randomly selected animals per genotype and time point for each experiment. One-way Anova with Tukey's multiple comparison test was used for statistical significance testing. p-value < 0.0001****, error bars represent standard deviation.

Fly strains and fly maintenance. The *RpII215^{C4}* fly strain (RRID:BDSC_3663), which carries a single point mutation (R741H) in the gene encoding the *Drosophila* RNA polymerase II 215kD subunit (RBP1), was received from the Bloomington *Drosophila* Stock Center (Bloomington, Indiana, USA). Flies carrying the *RpII215^{C4}* allele²³³ are homozygous viable but show a reduced transcription elongation rate²⁰⁵. *RpII215^{C4}* mutants were backcrossed for 6 generations into the outbred *white Dahomey* (*wDah*) wild type strain. A PCR screening strategy was used to follow the *RpII215^{C4}* allele during backcrossing. Therefore, genomic DNA from individual flies was used as a template for a PCR reaction with primers SOL1064 (CCGGATCACTGCTGCATATTTGTT) and SOL1047 (CCGCGCGACTCAGGACCAT). The 582 bp PCR product was restricted with BspHI, which specifically cuts only in the *RpII215^{C4}* allele, resulting in two bands of 281 bp and 300 bp. At least 20 individual positive female flies were used for each backcrossing round. Long-lived insulin mutant flies, which lack three of the seven *Drosophila* insulin-like peptides, *dilp2-3,5* mutants (RRID:BDSC_30889)²³⁴ were also backcrossed into the *wDah* strain, which was used as wild type control in all fly experiments. Flies were maintained and experiments were conducted on 1,0 SY-A medium at 25°C and 65% humidity on a 12L:12D cycle²³⁴.

Fly lifespan assays. For lifespan assays, fly eggs of homozygous parental flies were collected during a 12 h time window and the same volume of embryos was transferred to each rearing bottle, ensuring standard larval density. Flies that eclosed during a 12 h time window were transferred to fresh bottles and were allowed to mate for 48h. Subsequently, flies were sorted under brief CO₂ anesthesia and transferred to vials. Flies were maintained at a density of 15 flies per vial and were transferred to fresh vials every two to three days and the number of dead flies was counted.

Lifespan data were recorded using Excel and were subjected to survival analysis (log rank test) and presented as survival curves.

Mouse maintenance and dietary restriction protocol. The DR study was performed in accordance with the recommendations and guideline of the Federation of the European Laboratory Animal Science Association (FELASA), with all protocols approved by the Landesamt für Natur, Umwelt und Verbraucherschutz, Nordrhein-Westfalen, Germany. For the mouse kidney, male *C57BL/6* mice were housed under identical SPF conditions in group cages (5 or fewer animals per cage) at a relative humidity of 50-60% and a 12 hour light and 12 hour dark rhythm. For dietary restriction vs control, 8 week old mice were used. Dietary restriction was applied for 4 weeks. Control mice received food and water ad libitum. Mice were sacrificed at 12 weeks. For comparison of young vs aged mice, 14 week and 96 week old mice were used. Food was obtained from ssniff (Art. V1534-703, Soest, Germany) and Special Diet Services (Witham, UK). The average amount of food consumed by a mouse was determined by daily weighing for a period of two weeks and was on average 4,3 g per day. DR was applied for 4 weeks by feeding 70% of the measured ad libitum amount of food. Water was provided ad libitum. Mice were weighed weekly to monitor weight loss. Neither increased mortality nor morbidity was observed during dietary restriction.

2.3.2 Biochemistry and molecular biology methods

RNA extraction for next-generation sequencing.

C. elegans: Wild-type N2 strain, alpha-amanitin resistant *ama-1(m322)* mutants and long-lived insR/IGF signaling mutants, *daf-2(e1370)* were sent for RNA-seq. For each genotype, more than 300 aged-matched adult worms at desired time points were collected in Trizol (Thermo Fisher Scientific, USA) in 3 biological replicates. Total RNA was extracted using RNeasy Mini kit (Qiagen, Hilden, Germany).

D. melanogaster: The RNA-seq data for brains of 30 days and 50 days old *dilp2-3,5* and *wDah* control flies have been published previously.²³⁵ 10 days and 50 days old *Rpl1215^{C4}* mutants and *wDah* control flies were snap frozen and fly heads were isolated by vortexing and sieving on dry ice. Total RNA from three biological samples per treatment group was prepared using Trizol Reagent according to the manufacturer's instructions, followed by DNase treatment with the TURBO DNA-free Kit (Thermo Fisher Scientific).

M. musculus: Mouse liver samples were isolated from 5, 16 and 27 months old ad libitum and DR animals, which corresponded to 2, 13 and 24 months of DR treatment, respectively. RNA was isolated by Trizol and DNase-treated. The RNA-seq data for 5 and 27 months old liver DR samples have been published previously²³⁶. RNeasy mini Kit and Trizol were used to isolate RNA from snap-frozen kidneys as per manufacturer's instructions. Hypothalamus tissue of long-lived insulin receptor substrate 1 (*IRS1^{-/-}*) knock out mice²³⁷ and *C57BL/6* black control animals was dissected manually at the age of 27 months. RNA was isolated by Trizol with subsequent DNase treatment. For blood samples globin RNA was removed using GLOBINclear™ Kit, mouse/rat/human, for globin mRNA depletion.

Human whole blood sample RNA extraction. After storage at -80°C for at least 24 h RNA extraction was performed by usage of PAXgene Blood RNA Kit (Quiagen, Hilden, Germany) according to the manufacturer's protocol. The study was operated in accordance with the Declaration of Helsinki and the good clinical practice guidelines by the International Conference on Harmonization. All patients provided informed consent and approval of each study protocol was obtained from the local institutional review board (Ethics committee of the University of Cologne, Cologne, Germany; (17-362, 2018-01-17).)

Total RNA and nascent RNA sequencing. From 1 µg input of total RNA, ribosomal RNA was removed using the Ribo-Zero Human/Mouse/Rat kit (Illumina). Sequencing libraries were generated according to the TruSeq stranded total RNA (Illumina) protocol. To generate the final cDNA library, products were purified and amplified by PCR for 15 cycles. After validation and quantification of the library on an Agilent 2100 Bioanalyzer, equimolar amounts of libraries were pooled. Pools of 5 libraries were sequenced per lane on an Illumina HiSeq 4000 sequencer. For a description of all the RNA-seq datasets used in this study see Supplementary Table 2.2. The same protocol was used to sequence cDNA libraries from human cell "factory" RNA, which was isolated as described previously²³⁸.

4sU-DRB labeling and TUC-conversion.

First, transcription was reversibly inhibited by 6-dichlorobenzimidazole 1-β-d-ribofuranoside (DRB) in order to achieve accumulation of RNA polymerase II at the transcription start sites and synchronized transcriptional elongation initiation upon DRB removal. Simultaneously with the DRB removal, cells were pulsed for different time points with the Uridine-analogue 4-thiouridine (4sU) in order to enrich for newly synthesized transcripts. Finally, total RNA was isolated per each time point and the RNA polymerase II speed was determined by calculating the 4sU nucleotides added to the nascent transcript per time point. To estimate RNA polymerase II speed change in aging cells, human fetal lung fibroblasts (IMR90) in proliferating and in senescent state were treated using this experimental procedure.

In order to select the time-points to be used in the experiment, validate the DRB treatment and removal and check the enrichment efficiency of 4sU, a control experiment was set according to the protocol of Fuchs *et al.* 2015⁷⁷. Two million proliferating cells (passage 14) were treated with 100 µM DRB (Merck, D1916) in their medium for three hours at 37 °C and, upon DRB removal, they were pulsed with 1 mM 4sU (Sigma-Aldrich, T4509) for 0 min, 5 min, 15 min, 30 min, 45 min, 60 min, 90 min and 120 min. Immediately after the completion of each time point, cells were lysed in TRIzol (ThermoFisher, 15596018) and RNA was isolated with the Direct-Zol RNA mini-prep kit (ZymoResearch, R2052). To validate DRB treatment, qRT-PCR was performed in cDNA from all time points using the primers designed by Fuchs *et al.* 2015⁷⁷ in proximal and distant introns of the OPA-1 gene. Furthermore, to estimate 4sU enrichment, the RNA collected in each time point was biotinylated using the EZ-Link biotin HPDP kit (ThermoFisher, 21341) and biotinylated RNA was enriched with streptavidin-coated beads (DYNAL™ Dynabeads™ M-280 Streptavidin, ThermoFisher 11205D). qRT-PCR evaluation was performed also with the primers suggested by Fuchs *et al.* 2015⁷⁷ against TTC-17 nascent and mature mRNA and 18S rRNA.

For the actual experiment, we performed the Thiouridine-to-Cytidine Conversion Sequencing (TUC-Seq) protocol developed by Lusser *et al.* (2020)²³⁹ in order to detect the 4sU labeled transcripts in different time points. In this method, the thiol group of 4sU is quantitatively converted to cytidine via oxidation by OsO₄ in aqueous NH₄Cl solution. The OsO₄-treated RNA samples are submitted to RNA-sequencing to quantify labeled and unlabelled transcripts and define the number of reads containing Uridine-to-Cytidine conversions. To this aspect, nine million proliferating (passage 9) cells and nine million cells that had entered senescence (passage 35) were treated with 100 μM DRB for three hours at 37 °C. Immediately after DRB removal, cells were pulsed with 1 mM 4sU for 0 min, 5 min, 15 min, 30 min and 45 min. RNA was isolated manually according to the TRIzol protocol and treated with 40 Units DNase I (ZymoResearch, E1010) for 20 min at room temperature. RNA was purified with the RNA Clean & Concentrator-25 kit (ZymoResearch, R1018) and quantified using a NanoDrop spectrophotometer. For the TUC-conversion, 10 μg of 4sU labeled RNA was treated with 1.43 mM OsO₄ (Merck, 251755) in 180 mM NH₄Cl (Merck, 09718) solution pH 8.88 for 3 hours at 40 °C as described in Lusser *et al.* (2020)²³⁹. Subsequent sample concentration and purification were also done according to this protocol. 4sU-labeled and OsO₄-treated RNA samples derived from proliferating and senescent IMR90s in all five time-points were subjected to RNA-sequencing. As a negative control for the TUC-conversion, we used a mixture 1:1 of 4sU-labeled but not OsO₄-treated samples from the time-points 30 min and 45 min. The RNA-sequencing was performed in two biological replicates per condition.

³⁵S-methionine/³⁵S-cysteine incorporation to measure translation rates in *Drosophila*.

Ex-vivo incorporation of radio-labeled amino acids in fly heads was performed as previously described²⁴⁰. Briefly, 25 heads of each young (10 days) and old (50 days) wDah control and Rpl1215^{C4} mutant animals were dissected in replicates of 5 and collected in DMEM (#41965-047, Gibco) without supplements, at room temperature. For labeling, DMEM was replaced with methionine and cysteine free DMEM (#21-013-24, Gibco), supplemented with ³⁵S-labeled methionine and cysteine (#NEG772, Perkin-Elmer). Samples were incubated for 60 min at room temperature on a shaking platform, then washed with ice cold PBS and lysed in RIPA buffer (150 mM sodium chloride, 1.0% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris, pH 8.0) using a pestle gun (VWR, Germany). Lysates were centrifuged at 13.000 rpm and 4 °C for 10 min and protein was precipitated by adding 1 volume of 20% TCA, incubating for 15 min on ice and centrifugation at 13.000 rpm, 4 °C for 15 min. The pellet was washed twice in acetone and resuspended in 200 μl of 4 M guanine-HCl. 100μl of the sample was added to 10ml scintillation fluid (Ultima Gold, Perkin-Elmer) and counted for 5 min per sample in a scintillation counter (Perkin-Elmer). Protein determination was done in duplicates (25μl each) per sample using the Pierce BCA assay kit (Thermo Fisher Scientific). Scintillation counts were normalized to total protein content.

MNase-seq sample preparation.

Mononucleosomal DNA from proliferating and senescent IMR-90 cells (from two donors) were prepared and sequenced on an Illumina HiSeq4000 platform as previously described²⁴¹. For fly heads, a MNase digestion assay was performed using the EZ nucleosomal DNA prep kit, as per the manufacturer's guidelines (Zymo Research). Briefly, 25 snap frozen heads were lysed in nuclei prep buffer, and incubated on ice for 5min. Cuticle fragments were then removed via centrifugation

(30sec 50xg). Nuclei were pelleted (5min 500xg) and washed twice in digestion buffer and resuspended in 100µl of digestion buffer. Nucleosome footprints were then digested using 0.05U of MNase (Zymo Research). Samples were taken at 0, 2, 3 and 5 min or 10 min for prolonged digestion, and immediately stopped in MN stop buffer (Zymo Research). Samples were isolated using Zymo Spin IIC columns. Nucleosome footprints (1:10 dilution) were visualized by TapeStation using High sensitivity D1000 ScreenTape (Agilent).

Western blotting.

Western blots were carried out on protein extracts of individual dissected tissues. Proteins were quantified using BCA (Pierce). Equal amounts were loaded on Any-KD pre-stained SDS-PAGE gels (Bio-Rad) and blotted according to standard protocols. Antibody dilutions varied depending on the antibody and are listed here: Histone 3 (1:1000), HP1 (DSHB) (1:500). Appropriate secondary antibodies conjugated to horseradish peroxidase were used at a dilution of 1:10000.

Inducible histones overexpression.

Doxycycline (Dox)-inducible expression of histones H3 and H4 in proliferating human fetal lung fibroblasts (IMR90) was achieved using the PiggyBac transposition system²⁴². The open reading frames of H3 and H4 were cloned in the Dox-inducible expression vector KA0717 (KA0717_pPB-hCMV*1-cHA-IRESVenus was a gift from Hans Schöler, Addgene plasmid #124168; <http://n2t.net/addgene:124168>; RRID:Addgene_124168) fused at their 3' end in frame to the sequence of the yellow fluorescent protein (YFP) mVenus²⁴³. After sequencing validation, each construct was co-transfected in IMR90s with the transactivator plasmid KA0637 (KA0637_pPBCAG-rtTAM2-IN was a gift from Hans Schöler, Addgene plasmid #124166; <http://n2t.net/addgene:124166>; RRID:Addgene_124166) and the Super piggyBac Transposase expression vector (SBI System Biosciences, PB200PA-1,) using the Lipofectamine™ LTX Reagent with PLUS™ Reagent (ThermoFisher Scientific, 15338100,) according to the manufacturer's instruction. In total 2.5 µg of the vectors KA0717, KA0637 and PB200PA-1 were used for each transfection in a 10:3:1 ratio. Stable transgene-positive cells were selected using 250 µg/ml G418 (resistance gene carried in KA0637) for 7 days. Emerging cells were induced for 24 h with 2.5 µg/ml doxycycline and then subjected to Fluorescent-Activated Cell Sorting (FACS) to select the ones expressing the mVenus (BD FACSAria™ II, BD Biosciences). H3 and H4 overexpression was verified by Western Blot with anti-H3 and anti-H4 antibodies (Abcam, ab1791 and ab10158 respectively).

All further assays were repeated in proliferating cells and cells at the senescence entry state. Senescence state was monitored by Beta-galactosidase staining²⁴⁴ in different passages (Cell Signaling Technology, Senescence β-Galactosidase Staining Kit, 9860). Immunofluorescence stainings (IF) to detect HMGB1, p21 and HMGB2 (Abcam, ab18256, ab184640 and ab67282 respectively) were performed as previously described²⁴⁵ and images were acquired in a widefield Leica DMi8 S with an HCX PL APO 63x/1.40 (Oil) objective. For MTT assays²⁴⁶ 6000 cells of each condition were seeded per well in a 96-well plate in four replicates, incubated for 4 h at 37 °C after the addition of 1 mM MTT (3-(4,5-Dimethylthiazol-2-yl)-2,5-Diphenyltetrazolium Bromide, ThermoFisher, M6494), treated with DMSO for 10 min at 37 °C and finally their absorbance was measured at 540 nM in an Infinite® 200 PRO plate reader (Tecan). For RNA-sequencing, one million cells of each condition were lysed in Trizol (ThermoFisher, 15596018) and RNA was isolated with the Direct-Zol RNA mini-prep kit (ZymoResearch, R2052).

2.3.3 Computational methods

RNA-seq alignments

Raw reads were trimmed with trimmomatic version 0.33²⁴⁷ using the following parameters 'ILLUMINACLIP:./Trimmomatic-0.33/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:45' for paired-end datasets and 'ILLUMINACLIP:./Trimmomatic-0.33/adapters/TruSeq3-SE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:45' for single-end datasets. Alignment was performed with STAR version 2.5.1b²⁴⁸ using the following parameters: '--outFilterType BySJout --outWigNorm RPM' on the genome version mm10, rn5, hg38, dm6, ce5 for *M. musculus*, *R. norvegicus*, *H. sapiens*, *D. melanogaster*, *C. elegans*, respectively. The bam files produced by STAR were used for all analyses, unless otherwise noted.

Definition of intronic regions.

All annotation files for this analysis were downloaded from the Ensembl website²⁴⁹ using genome version ce10 for *Caenorhabditis elegans*, mm10 for *Mus musculus*, hg38 for *Homo sapiens*, rn5 for *Rattus norvegicus*, and dm6 for *Drosophila melanogaster*. All of the following filtering steps were applied on the intronic ENSEMBL annotation files.

First, overlapping regions between introns and exons were removed in order to avoid confounding signals due to variation in splicing or transcription initiation and termination. Overlapping introns were merged to remove duplicated regions from the analysis. In the next step, STAR was used to detect splice junctions. These junctions were then compared to the intronic regions. Introns with at least 5 split reads bridging the intron (i.e. mapping to the flanking exons) per condition were kept for subsequent analyses. This step ensures a minimum expression level of the spliced transcript. When splice junctions were detected within introns, those introns were further subdivided accordingly. Introns with splice junction straddling were discarded. The above-mentioned steps were performed using the subtract and merge commands of Bedtools version 2.22.1.

After these filtering steps, the number of usable introns per sample varied between a few hundred ($n=546$, *C. elegans*, total RNA) to over ten thousand ($n=13,790$, *H. sapiens*, nascent-RNA-seq). These large differences resulted from different sequencing depths, sequencing quality (number of usable reads) and from the complexity of the genome (numbers and sizes of introns, number of alternative isoforms, etc.). In order to avoid artifacts due to the different numbers of introns used per sample, the same sets of introns were always contrasted for each comparison of different conditions (e.g. old *versus* young, treatment *versus* control). Note that certain comparisons were not possible for all species, due to variations in the experimental design. For instance, for mouse kidney only a single time point after lifespan intervention (dietary restriction DR, age 3 months) was available, which prevented a comparison of old versus young DR mice, but allowed comparison with *ad libitum* fed mice at the young age.

Transcriptional elongation speed based on intronic read distribution.

In order to calculate Pol-II speeds, we used RNA-seq data obtained from total RNA and nascent RNA enrichment. In contrast to the widely used polyA enrichment method, which primarily captures mature, spliced mRNAs and is therefore not suitable to estimate Pol-II speeds based on intronic

reads, these methods yield sufficient intronic coverage to quantify elongation rates. To analyze the distribution of intronic reads between conditions, we fitted the read gradient (slope) across the length of each of selected introns (5'→3'; see above for the filtering criteria). The read gradient was calculated from the bedgraph files produced by STAR.

In order to transform slopes to Pol-II elongation speed the following formalism was used. Let us assume an intron of length L . Let us also assume that at steady state a constant number of polymerases is initiating and the same number of polymerases is terminating at the end of the intron; in other words, we assume that premature termination inside the intron can be ignored. Polymerases are progressing at a common speed of k [bp/min]. The average time that it takes a polymerase to traverse the whole intron is hence

$$\Delta t = \frac{L}{k}$$

Transcription is initiated at a rate of n polymerases per unit time [1/min]. Hence, the number of polymerases N initiating during Δt is:

$$N = \Delta t \times n$$

The slope s is the number of transcripts after the distance L minus the number of transcripts at the beginning divided by the length of the intron:

$$s = \frac{0-N}{L} = \frac{-\Delta t \times n}{L} = \frac{\frac{-L}{k} \cdot n}{L} = \frac{-n}{k}$$

and thus, the speed k can be computed from the slope as:

$$k = \frac{-n}{s}$$

Hence, slope and speed are inversely related and the speed depends also on the initiation rate (i.e. the expression rate). However, we observed empirically only a small dependency between expression and slope (Supplementary Table 2.1).

Transcript counts were estimated using Kallisto version 0.42.5²⁵⁰ for each sample. To determine differentially expressed genes we used DESeq2 version 1.8.2²⁵¹ with RUVr normalization version 1.6.2²⁵². For the differential analysis of transcriptional elongation regulators, we downloaded the list of positive and negative regulators from the GSEA/MSigDB²⁵³. Gene ontology (GO) term enrichment analysis of differentially expressed genes or genes with increased RNA-Pol-II elongation speed was carried out using TopGO version 2.20.0. For GO enrichment analysis of differentially expressed genes, we identified 4784 genes as evolutionarily conserved from each species of our study to humans: genes were either direct orthologues (one2one) or fusion genes (one2many) of *H. sapiens* were retrieved from ENSEMBL database using biomaRt 2.24.1²⁵⁴. Using our 4784 genes evolutionarily conserved, we further divided into consistently up-regulated or down-regulated genes across species during aging or 'aging intervention' (as target set for GO

enrichment: aging up-regulated: 92 genes; aging down-regulated: 71; ‘aging intervention’ up-regulated: 164 genes; ‘aging intervention’ down-regulated: 473 genes; as background set 4784 orthologue genes between *R. norvegicus*, *M. musculus*, *D. melanogaster*, *C. elegans*, and *H. sapiens*). For GO enrichment analysis of genes harboring increasing Pol-II speed, we used as target set the top 200 or 300 genes with an increase in Pol-II speed change for each species. Quantification of transcript abundance for ITPR1 and AGO3 was obtained by using StringTie²⁵⁵. For circular RNA, we aligned the reads using STAR version 2.5.1b²⁴⁸ with the following parameters: ‘--chimSegmentMin 15 --outSJfilterOverhangMin 15 15 15 15 --alignSJoverhangMin 15 --alignSJDBoverhangMin 15 --seedSearchStartLmax 30 --outFilterMultimapNmax 20 --outFilterScoreMin 1 --outFilterMatchNmin 1 --outFilterMismatchNmax 2 --chimScoreMin 15 --chimScoreSeparation 10 --chimJunctionOverhangMin 15’. We then extracted back spliced reads from the STAR chimeric output file and normalized the number of back spliced reads by the sum of back spliced (BS_i) and spliced reads from linear transcripts ($S1_i$, $S2_i$) for an exon i^A :

$$CircRatio_i = \frac{BS_i}{BS_i + \frac{S1_i + S2_i}{2}} * 100$$

Here, $S1_i$ refers to the number of linearly spliced reads at the 5’ end of the exon and $S2_i$ refers to the respective number of reads at the 3’ end of the exon. Thus, this score quantifies the percent of transcripts from this locus that resulted in circular RNA. Finally, we quantified the significance of the average change in circular RNA formation between two conditions using the Wilcoxon rank test.

To validate our estimates of Pol-II speeds we compared our data with experimental values estimated via GRO-seq⁸³ and tiling microarray data⁵⁶. There was a significant correlation (GRO-seq: $R=0.38$, $p\text{-value}=4e-5$, compared to time point 25-50 min see Jonkers et al.⁸³; Tiling array; $R=0.99$, $p\text{-value}<2.6e-16$, data not shown) between our data and experimentally measured transcriptional elongation values. We noted that our Pol-II speed estimates for different introns of the same gene were more similar than Pol-II speed estimates for random pairs of introns, implying that gene-specific factors or local chromatin structure influence Pol-II speed.

4sU-DRB elongation rate calculation.

The estimation of transcription elongation speed using RNA labeling was based on the measurement of nucleotides added per time unit in a newly synthesized nascent transcript.

Detection of labeled transcripts was performed based on the Lusser protocol, modified for Illumina RNA-seq:

1. FASTQ files were aligned to the genome using STAR to produce BAM files.
2. Sam2tsv was then used to identify single nucleotide mismatches.
3. A custom R script was used to count the number of A-G or T-C mismatches per read.

Only read-pairs with at least 3 A-G or T-C mismatches were assumed to be 4sU-labeled and thus retained for subsequent analyses. Because the 5 min samples contained a very low number of reads with conversions, they were discarded from the rest of the analysis. We employed two approaches for estimating the elongation rate per gene from the 4sU-labeling data. For the first

approach, we tracked the progress of RNA-Pol-II complexes constructing single gene coverage profiles using 4sU-labeled reads. Progression was determined by picking the 99th percentile of gene body coverage in each sample to determine the front of elongating RNA polymerases. (We did not use the last converted read to determine the front, because this measure would be too sensitive to noise in the data.) Elongation rates were calculated by fitting a linear model on the front positions of Pol II in 0, 15, 30 and 45 min in the first 100 kb of each gene. To determine elongation rates with greater accuracy, we filtered out genes with a length of less than 100 kb, since short genes can be fully transcribed in less than 45 min or even 30 min. This first approach of estimating RNA-Pol-II speeds is characterized by high accuracy, but is limited to genes longer than 100 kb. The data in Panel e of Figure 2.1 is based on this approach.

The direct comparison of the 4sU data to the approach using read-coverage slopes in introns required a large set of genes for which RNA-Pol-II speed could be measured using both assays. In order to maximize this gene set we devised a second alternative approach for deriving speed from 4sU-labeling data that is applicable to shorter genes. For this second approach we measured the front position of the polymerase in the same way as before (using the 99th percentile) but across the whole gene. For genes 30 kb to 100kb long, we calculated the elongation rate from the difference in the front positions of the polymerase at 15 mins and 30 mins and divided this distance by 15 minutes in order to obtain speed measures per minute. For genes more than 100kb long, we calculate the elongation rate from the difference in the positions of the polymerase at 30 mins and 45 mins divided by 15 minutes. This second speed measure is less accurate than the first one, because it uses only two time points per gene; however, it enables estimating speed for genes shorter than 100 kb. The data in Panel d of Figure 2.1 is based on this second measure. Note that both measures confirmed the increase in average RNA-Pol-II elongation speed from proliferating to senescent IMR90 cells.

Mismatch detection.

Mismatch detection was performed using the tool rnaseqmut (<https://github.com/davidliwei/rnaseqmut>), which detects mutations from the NM tag of BAM files. To avoid detection of RNA editing or DNA damage-based events we only considered genomic positions with only 1 mismatch detected (i.e. occurring in only one single read). Reads with indels were excluded and only mismatches with a distance of more than 4 from the beginning and the end of the read were considered. A coverage level filter was applied so that only bases covered by at least 100 reads were kept. A substantial number of mismatches may result from technical sequencing errors. However, since young and old samples were always handled together in the same batch, we can exclude that consistent differences in the number of mismatches are due to technical biases. The fraction of RNA editing events is generally relatively low and not expected to globally increase with age²⁵⁶.

MNase-seq analysis.

We used nucleR²⁵⁷ with default parameters to calculate peak sharpness as a combination of peak width and peak height. Peak 'width' was quantified as the standard deviation around the peak center and peak 'height' was quantified as the number of reads covering each peak²⁵⁷ and the distance between peak summits. Intron and exon annotations were downloaded from UCSC table

utilities²⁴⁹ and filtered as described in Definition of intronic regions. Nucleosome density (Figure 5a) is defined as the number of nucleosome peaks found within an exon or an intron divided by the length of the exon or intron.

2.4 Discussion

We found a consistent increase in average intronic Pol-II elongation speed with age across four animal models, two human cell lines and human blood, and could revert this trend by employing lifespan-extending treatments. We also documented aging-related changes in splicing and transcript quality, such as elevated formation of circular RNAs and increased numbers of mismatches with genome sequences, which likely contribute to age-associated phenotypes. Further, we observed a consistent increase in the ratios of spliced to unspliced transcripts (splicing efficiency) with age across species (Fig. 2.3a), which has been reported to be a result of increased elongation speed¹³². However, we cannot exclude the possibility that this increase resulted from changes in RNA half-lives. Although average speed changes were predominantly significant, they remained small in absolute terms. This is expected, as drastic, genome-wide changes of RNA biosynthesis would quickly be detrimental for cellular functions and likely lead to early death. Instead, what we monitored here is a gradual reduction of cellular fitness characteristic for normal aging. Critically, we were able to increase lifespan in two species by decelerating Pol-II. Thus, despite being small in magnitude and stochastically emerging in tissues or cell populations, these effects are clearly relevant for organismal lifespan.

Genes exhibiting accelerated Pol-II elongation were not enriched for specific cellular processes, indicating that speed increase is probably not a deterministically cell-regulated response, but rather a spontaneous age-associated defect. Yet, the genes affected were not completely random, since we observed consistent changes across replicates for a subset of introns. Thus, there must be location-specific factors influencing which genomic regions are more prone to Pol-II speed increase and which not. This observation is consistent with earlier findings and our data, indicating that chromatin structure may causally contribute to age-associated Pol-II speed increase. Although we still lack a complete understanding of the molecular events driving Pol-II speed increase, our findings indicate that aging-associated changes in chromatin structure play an important role.

Our work establishes Pol-II elongation speed as an important contributor to molecular and physiological traits with implications beyond aging. Misregulation of transcriptional elongation reduces cellular and organismal fitness and may therefore contribute to disease phenotypes^{228,258,259}. Taken together, the data presented here reveal a new molecular mechanism contributing to aging and serve as a means for assessing the fidelity of the cellular machinery during aging and disease.

2.5 Contributions

The work described in this chapter is available in the following preprint:

Debès, C. et al. Aging-associated changes in transcriptional elongation influence metazoan longevity. Preprint at <https://doi.org/10.1101/719864> (2022).

All the experimental work was done by the other authors of the paper, from the groups of Linda Partridge, Argyris Papantonis, Adam Antebi, Roman-Ulrich Müller, Bernard Schermer and Thomas Benzing. Isabell Brusius performed the LeafCutter splicing analysis (Figure 2.3a). Cedric Debès wrote the code for the slope estimation method and circular RNA fraction calculation from RNA-seq reads. He also calculated the transcriptional mismatch level changes (Figure 2.3c), analyzed gene expression changes (Supplementary Figures 2.5, 2.6 and 2.7) and created Supplementary Figure 2.10. I performed the rest of the bioinformatics analyses, including the definition of intronic regions, Pol II speed calculation, 4suDRB-seq analysis, splicing efficiency estimation, circular RNA detection and MNase-seq analysis. I wrote the manuscript together with Andreas Beyer and Cedric Debès.

Chapter 3. Age-related changes in transcriptional fidelity across tissues

3.1 Introduction

Protein synthesis is an error-prone process²⁶⁰. Errors in the amino acid sequence can cause defects in cellular fitness²⁶¹ and lead to the emergence of diseases^{262,263}. Therefore, precision in the mechanisms that maintain the faithful expression of our genetic code, like DNA replication, transcription and translation, is vital. There have been numerous studies focused on the mutations caused by DNA replication²⁶⁴, as they are heritable, and translation^{163,265}, since the error rate of translation is normally an order of magnitude higher than the one of transcription. Transcriptional infidelity²⁶⁶ is less well studied, since there are technical limitations that inhibit the effort. However, errors in transcription can have worse consequences than those in translation, since a single error can be massively amplified²⁶⁷. As a consequence, if the error changes the function of a crucial or long-lived protein, it can have a very significant effect on the fate of the cell or even the organism. Previous studies have shown that an increase in the error rate of transcription can have very deleterious effects on cellular homeostasis and cell fate in general^{5,11} and that transcriptional infidelity can contribute to aging²⁶⁸ and aging-related diseases.

Precisely measuring the error rate of transcription of protein coding genes would therefore be important to further our understanding of the causes and consequences of transcriptional infidelity and its relation to aging. *In vitro* assays^{146,163} have estimated the error rate of Pol II to be around 1×10^{-5} . However, in living organisms there are several factors (including but not limited to transcription factors, DNA damage, chromatin structure and repair mechanisms) that may influence the real error rate of transcription. Additionally, *in vitro* assays are limited in the spectrum of errors and sequence contexts they can monitor²⁶⁹. As a result, in order to properly understand the effects of Pol II errors, it seems it is necessary to study transcriptional fidelity *in vivo*.

Next generation sequencing technologies seem ideal for this task. After all, RNA-seq has already been used to great effect to study DNA mutations and there is already a huge variety of published datasets, which could be mined. Single-cell RNA sequencing data would be even more interesting, since they would provide the ability to distinguish between cell types and thus study the heterogeneity of transcriptional fidelity in tissues. Nevertheless, there are significant limitations regarding the detection of mismatches in RNA sequences. This is partially due to the fact that the step of reverse transcription has a much higher error rate compared to RNA synthesis, effectively masking transcriptional mistakes²⁷⁰. This issue is compounded by the fact that there are two other processes that can artificially inflate the calculated error rate: PCR amplification and sequencing. Both of those steps can cause significant artifacts¹⁷⁴. While there are sequencing approaches that significantly increase error rate estimation accuracy, like NET-Seq¹⁸⁰ and CirSeq¹⁷⁶⁻¹⁷⁸, they are technically challenging¹⁷⁹ and they cannot be used for standard published bulk or single-cell RNA-seq data.

In order to overcome these obstacles, we have developed scErrorRate, the first tool that can perform *in vivo* transcriptional error rate estimation from scRNAseq data. Using barcoded scRNAseq alignment files, a list of cell barcodes and a GTF file of the transcriptome, scErrorRate leverages the presence of UMIs (unique molecular identifiers) in single-cell data to identify sequence mismatches on a single-molecule level. UMIs are used to create a consensus sequence for the mRNA molecules from the barcoded reads and the molecular sequences covering the same part of the transcriptome are compared and the error rate for each cell is calculated.

3.2 Results

The detection of mismatches between aligned reads and the reference genome is a simple operation. Determining which of them are caused by Pol II infidelity is a more complicated affair, since transcriptional errors account for a small percentage of the detected mismatches. The rest of them are caused by (1) DNA polymorphisms, (2) sequencing mistakes (whether reverse transcription errors, PCR errors or alignment errors), (3) RNA editing and (4) RNA.

To enrich real transcriptional mistakes, scErrorRate generates consensus reads for every RNA molecule from single-cell RNA sequencing data. This is based on the link between the reads and the original mRNA derived from the unique molecular identifier of each read, a randomized nucleotide sequence that is distinct for each RNA molecule. UMIs have been widely used to generate more accurate estimations of gene expression in single-cell experiments, since they facilitate the detection and removal of PCR duplicates^{182,271}. Given sufficient reading depth, UMI tags can be leveraged to create a consensus sequence for each molecular identifier, eliminating artifacts caused by sequencing mistakes. This happens in two steps. First, a pileup table is created from the reads of every cell. Then, a proofreading step is applied on the pileup, creating a consensus sequence per cell per UMI per position. The RNA sequences covering the same part of the transcriptome are then compared to each other. After filtering out mismatches among the sequences that are likely to be caused by DNA polymorphisms and RNA editing, true mistakes can be detected with higher accuracy (see 3.3 for more detail). It is worth noting that the pipeline cannot distinguish between mistakes caused by reverse transcription and transcription errors. However, for relative comparisons, if the different samples were always handled together in the same batch, the possibility that consistent differences in the error frequency are due to technical biases during RT-PCR can be excluded.

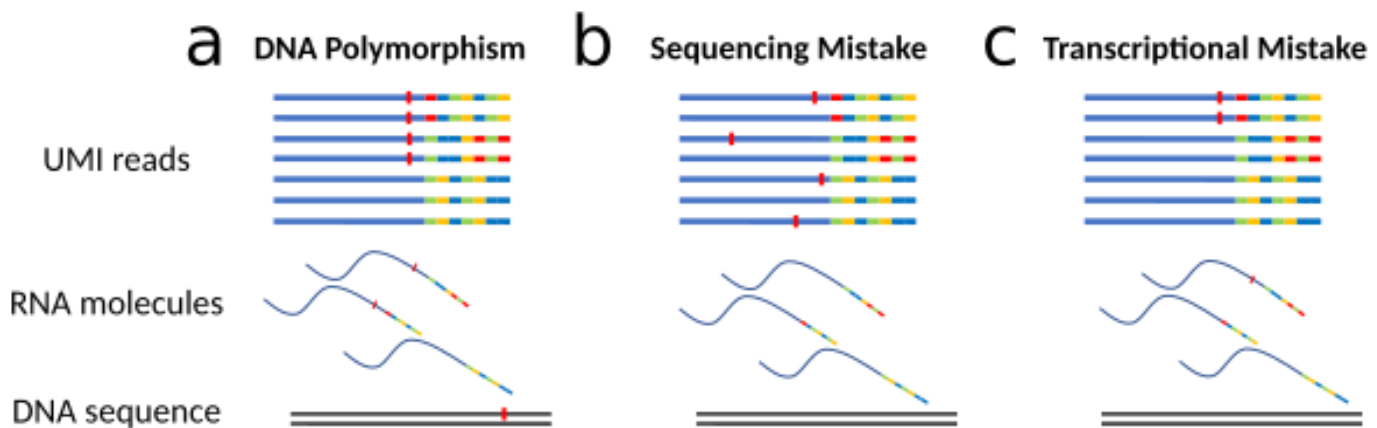


Figure 3.1: A visual representation of the idea behind the transcription error detection pipeline. The pipeline identifies transcriptional mistakes by leveraging the UMI information present in droplet-based single cell RNA-seq data to construct consensus mRNA sequences. (a) If there is a DNA polymorphism, either a single nucleotide polymorphism or an error in the DNA sequence, the error will be present in multiple transcripts (but not necessarily all of them, given the existence of heterozygous polymorphisms). If a

mismatch between consensus sequences appears more than one time, it is removed. **(b)** If there is a sequencing mistake, whether because of PCR amplification or because of the introduction of a base miscall from the Illumina machine, it is unlikely to be present in all reads that come from the same transcript. During the construction of the consensus sequence, these errors are thus eliminated. **(c)** If an error happens during transcription, it will be present in all the reads of the transcript, meaning the reads with the same UMI. It can be detected by comparing the consensus sequences of different UMIs covering the same genomic region.

ScErrorRate generates a cell-error matrix with the absolute counts of positions with errors (n_{err}), positions without errors (n_{con}) and the error rate ($ER = n_{err} / (n_{err} + n_{con})$) and a cell matrix containing the n_{er} , n_{con} and ER of every type of misincorporation error. Optionally, it can create tables with the exact positions of the epimutations and error coverage plots for specified regions or classes of genes, measure the frequency of specific motifs and generate an R object that can be incorporated in standard scRNAseq analysis and visualization. Strengths of scErrorRate, beyond the fact that it is the first tool to estimate RNA epimutations from scRNAseq data, include the uncomplicated format of its input and output, its explicit and user configurable capabilities for extraction of cell barcodes and quality filtering and its interoperability with Seurat²⁷². Its output can be used for a variety of downstream analyses.

We used publicly available data reporting single-cell transcriptomics measurements during aging to test the transcriptional error pipeline. Tabula Muris Senis²⁷³ or ‘Mouse Ageing Cell Atlas’ contains scRNAseq data from 23 tissues and organs across the lifespan of *Mus musculus*. We chose spleen samples from 1-month-old and 30-month-old mice, produced through droplet-based cell isolation paired with sequencing of the 3’ end of transcripts. The raw sequencing files were realigned and the cells were filtered to remove barcodes corresponding to empty droplets and doublets (Methods).

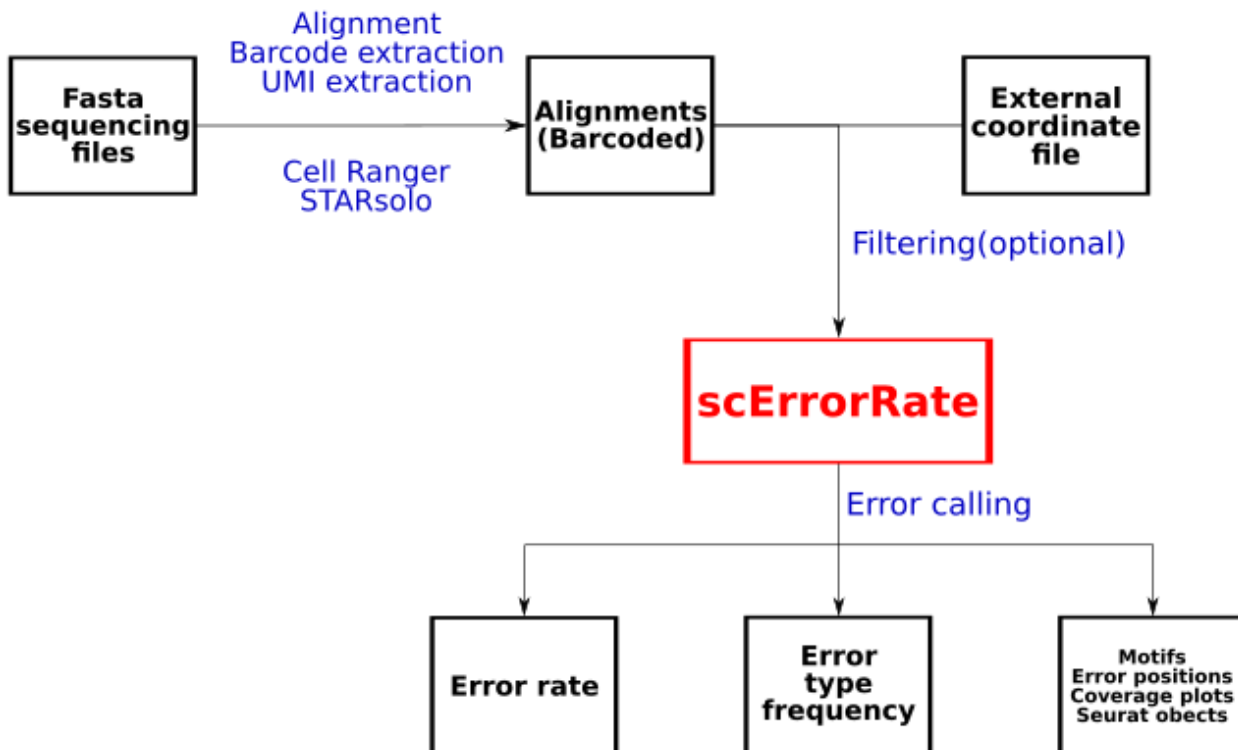


Figure 3.2: ScErrorRate workflow.

One possible caveat of scErrorRate is that it may be unable to differentiate between errors and RNA editing or other post-transcriptional modifications. The most common RNA editing in metazoans is adenosine-to-inosine (A-to-I) editing, in which the deamination of adenosine ribonucleotides to inosines is catalyzed by the adenosine deaminase acting on RNA (*ADAR*) family of genes^{274,275}. More than 99% of edits occur in Alu repeats and the majority are A-to-I, which appear as A-to-G transitions by RNA-seq. RNA editing mostly happens co-transcriptionally^{8,276}, but deamination of nucleobases in RNA can also be the result of spontaneous deamination and nitrosative stress²⁷⁷. To evaluate the pervasiveness of deamination caused by post-transcriptional RNA damage and RNA editing events, we generated a scRNAseq dataset with a novel nascent RNA-seq protocol. Usually, the capturing and barcoding of scRNAseq is performed through reverse transcription of their poly (A) tails²⁷⁸. To capture nascent RNA instead, the previously established “factory RNA-seq” protocol⁸⁴ was used to capture nuclei rich in nascently-transcribed RNA molecules from proliferating and senescent HUVEC cells. DNase digestion was used to remove chromatin that was not being actively transcribed. Polyadenylation of nascent transcripts was performed, followed by standard library preparation for the 10X Genomics platform. The files were aligned and filtered as described in the methods.

We explored an assortment of different scErrorRate applications on the different samples. For all samples, gene expression was estimated from the STARsolo output, and normalized and scaled with scTransform. For the published dataset, we used the same cell type annotation and clustering as in the original publications. Despite the fact that the original study used CellRanger to align the data, dimensional reduction and clustering clearly separated the dataset in the originally annotated cell types (Supplementary Figure 3.1). For the HUVEC nascent RNA-Seq data, we performed filtering as described in the method section.

For all samples, the number of detected mistakes per cell is significantly correlated with the number of UMIs and number of genes per cell, but not with the cellular proportion of mitochondrial reads (Supplementary Figure 3.2). The error rate itself is uncorrelated with all the previously mentioned cell metrics (Supplementary Figure 3.3). When we compare the proliferating cells to the senescent ones, we observe a statistically significant increase in cellular error rate (Figure 3.3a). Furthermore, the overall average error rate, calculated by dividing the number of mistakes in all the cells in a sample by the total coverage of that sample, is higher in the senescent sample compared to the proliferating one (Figure 3.3b). Similarly, in the spleen dataset, we observe that the error rate increases with age in all old replicates.

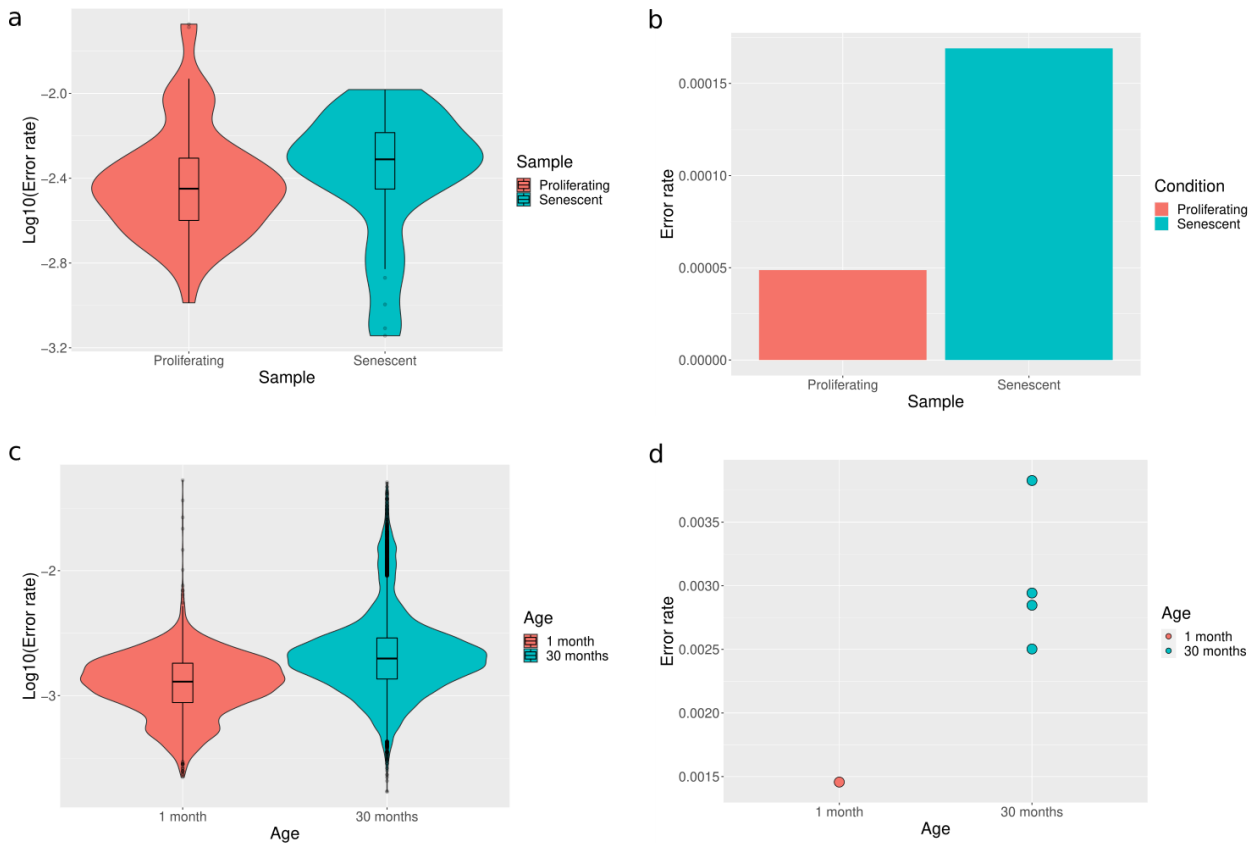


Figure 3.3: Changes in transcription error rate with senescence and aging. (a) Distributions of error rates in HUVEC cells. The difference in error rates between proliferating and senescent cells is statistically significant (unpaired Wilcoxon test, $p\text{-value} = 2.13\text{e-}10$). (b) Mean error rate of transcription in proliferating and senescent human cells. (c) Distributions of error rates in spleen cells. The difference in error rates between young and old cells is statistically significant (unpaired Wilcoxon test, $p\text{-value} < 2.2\text{e-}16$). (d) Mean error rate of transcription in young and old spleen samples.

Using a next generation sequencing method for error detection instead of a gene reporter assay allows genome-wide detection of transcription errors and localization to specific transcripts. Indeed, the detected errors span the entire human genome (Supplementary Figure 3.4). The number of errors per chromosome does not correlate with chromosomal length, but it does correlate with the chromosomal coverage of the consensus sequences. One issue with using `scErrorRate` on 10x sequencing data is that errors were mostly detected at the 3' end of the gene. This limits the potential to derive conclusions about which transcript regions are most affected (Supplementary Figure 3.5). We then calculated the error rate per gene. Within genes, error rate is not significantly correlated with gene expression (Figure 3.4a,b). Aging significantly increases the average error rate in coding genes in both the HUVEC and spleen datasets.

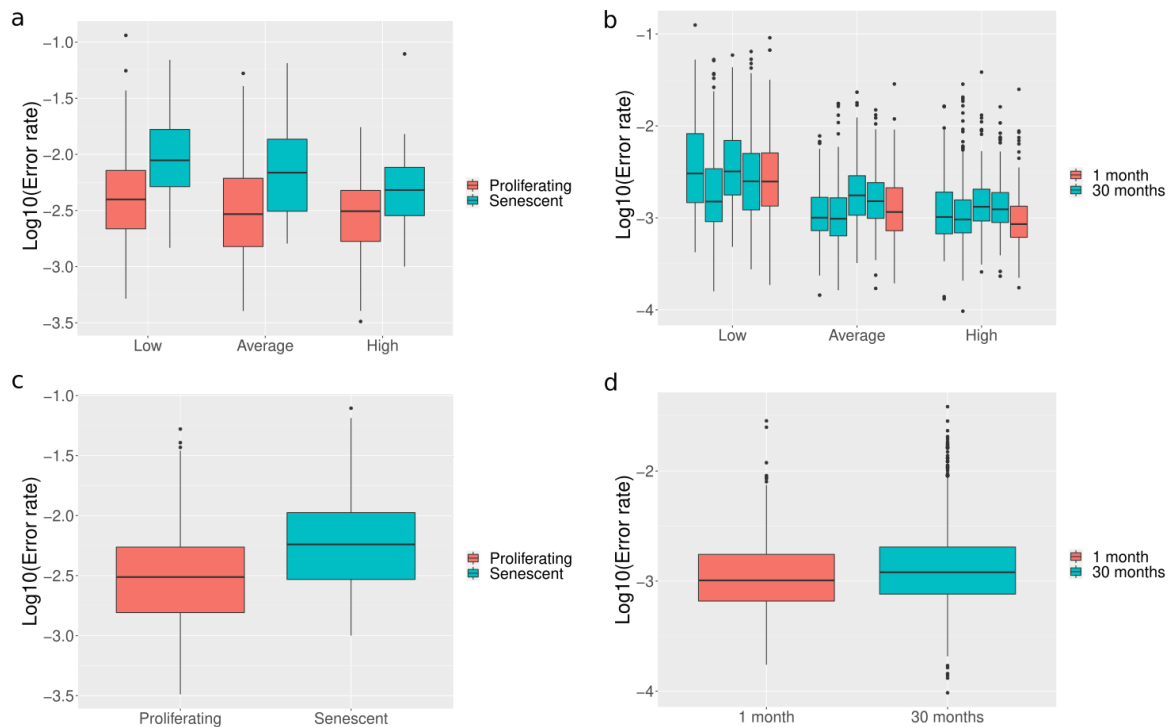


Figure 3.4: Genewise error rate. Genes were binned according to expression per sample in three categories of equal size in (a) HUVEC and (b) spleen cells. Senescence and aging significantly increase the average gene error rate in the (c) HUVEC and (d) spleen dataset (Wilcoxon rank sum test in HUVEC: p -value= $1.467e-14$, Wilcoxon rank sum test in spleen= 0.0004197). Genes were included if there was at least one mistake detected in one of the samples.

Finally, in order to better understand the molecular mechanisms that affect Pol II error rate, we investigated its error spectrum in greater detail. Specifically, we examined whether the primary sequence of the human genome affects mistake incorporation by computing the error rate on all nucleotides. Every kind of misincorporation increases its frequency with senescence in both samples (Figure 3.5). The error spectrum in the spleen matches what has been previously reported; transitions occur more often than transversions¹⁷⁵ and the most common error is C-to-U. This has been previously reported in human cells, yeast and worms^{175,179,279}. However, the error spectrum in HUVEC cells is substantially different; U-to-G is the most common reported mismatch, especially in senescent cells. It is possible that a fraction of the C-to-U RNA mutations observed in our polyA data is a result of deamination, either enzymatic or spontaneous, instead of transcriptional misincorporations.

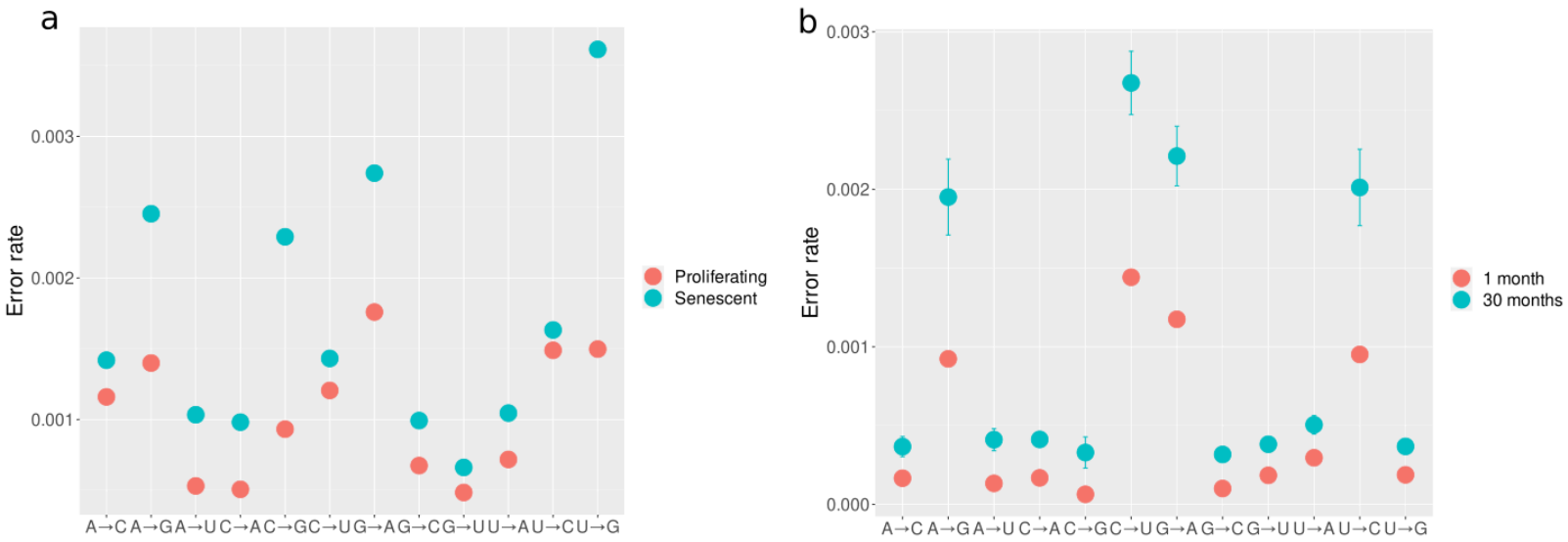


Figure 3.5: Spectrum of error types. The frequency of each RNA error type was determined by the number of errors observed over the total number of observations (number of positions in the consensus sequences) of the wild-type nucleotide in the (a) HUVEC and (b) spleen datasets.

3.3 Materials and methods

Pre-existing single-cell RNA-seq dataset

Spleen fasta sequencing data from the Tabula Muris Senis dataset were downloaded from a public AWS S3 bucket (<https://registry.opendata.aws/tabula-muris-senis/>). The mice and organ collection, tissue preparation, and the specifics of the 10x Genomics sequencing protocol are described in detail in the original study. Briefly, single cells were captured in droplet emulsions using the GemCode Single-Cell Instrument and scRNA-seq libraries were constructed as per the 10x Genomics protocol using GemCode Single-Cell 3' Gel Bead and Library V2 Kit. Libraries were sequenced on the NovaSeq 6000 Sequencing System (Illumina).

Single-cell nascent RNA sequencing

Proliferating and senescent HUVEC cells were washed once in an isotonic near-physiological buffer (PB) that maintains the cells' transcriptional activity and subjected immediately to the first steps of the "factory RNA-seq" protocol²³⁸. In more detail, cell nuclei are gently isolated using PB+0.4% NP-40, DNase I-treated at 33°C for 25 min to detach most chromatin, pelleted and washed once in ice-cold PB, before polyadenylation of nascent RNA²⁸⁰. Next, ~2,500 cells from each state were loaded onto the Chromium 10X Genomics platform for encapsulation in oil droplets and generation of barcoded cDNA libraries from individual nuclei as per manufacturer's instructions.

Data pre-processing

The fasta files were aligned to the mouse (mm10) and human (hg38) genome with the STARsolo²⁸¹ module of STAR v.2.7.8a, using the following parameters: `--soloType Droplet --soloUMIfiltering MultiGeneUMI --soloCBmatchWLtype 1MM_multi_pseudocounts --outSAMtype BAM SortedByCoordinate --soloFeatures GeneFull --outSAMattributes CB UB GX GN --readFilesCommand zcat`. StarSolo was also used for read-to-gene assignment, cell barcode demultiplexing with knee filtering, error correction, and unique molecular identifier (UMI) collapsing.

Gene expression was estimated from the filtered read count matrix output by STARsolo, and normalized and scaled using the `sctransform`²⁸² function of Seurat²⁷². For cell type determination of the spleen data, we used the `cell_ontology_class` column within the metadata table contained in the h5ad files of Tabula Muris Senis (<https://doi.org/10.6084/m9.figshare.8273102.v2>).

Pipeline data input

In order to run the pipeline, three files are necessary: the position-sorted, UMI-based scRNAseq alignment file, a file containing a list of error-corrected cell barcodes and a Gene Transfer Format (GTF) file of the transcriptome. The alignment files can be in either SAM or BAM format. The UMIs must be either under a specified read tag (CB by default) or in the read name. In the second case, a regular expression must be provided to extract the barcode from the name. SAMtools²⁸³ convert the files into BAM if they are SAM and index them. The cell barcode file should be a one-column text file containing the cell barcodes of interest. The pipeline can technically be run without a cell

barcode file, but it is highly recommended to first find the cell barcodes that correspond to real cells. The vast majority of barcodes contain very few reads and are produced from empty droplets, especially in 10X sequencing. If alignment is performed with CellRanger¹⁸⁵ or STARsolo²⁸¹, a filtered list of barcodes, corrected for sequencing errors, is produced by default. The GTF file has to be from the same genome annotation and genome version used for alignment for accurate results.

Pipeline workflow

The R parts of the pipeline were developed in R 4.2.2.

1. Splitting the BAM file

Specified read tags or the read name are used to parse the alignment file to create separate bam files for individual cells. This process is performed with Sinto 0.9.0 (<https://github.com/timoast/sinto>). The files are then indexed with SAMtools²⁸³.

2. Pileup

A pileup consists of a summary of the reads overlapping each genomic position, differentiating on nucleotide, strand and position within the read. Since every read is associated with a specific UMI, every nucleotide can also be tagged with the UMI it corresponds to and thus the molecule from which it was derived. This step goes through all the reads and generates a dataframe with the nucleotides in each position (chromosome, location and strand), the UMI sequence, the cell barcode, the sequence quality score and the mapping quality score. Reads from mitochondrial genes are excluded from the analysis, since each cell contains between 1000 and 10,000 copies of the mitochondrial genome²⁸⁴. Ambiguous nucleotides, deletions and insertions are removed by default. After completing the pileup, the pipeline generates an .rds file. This file can be used to try different parameters for the consensus sequence and error calling steps.

3. Consensus sequence calling

This step consists of collapsing all reads tagged with the same UMI into a single read. The purpose of this step is to reduce the number of false positive calls, either from PCR amplification or sequencing errors. All reads tagged with the same UMI should be produced by the same mRNA molecule and thus reads from the same genomic position should have the same nucleotide sequence. If there are mismatches between reads from the same UMI, they can be attributed to sequencing or amplification artifacts.

By default, positions with less than 5 reads are filtered out, since it is impossible to determine whether a mismatch is an artifact if the coverage is too low. If the sequencing depth per cell is very high, increasing this threshold is possible. To calculate the consensus read per UMI, there are two parameters that are used: the frequency of the nucleotide in each position and the quality of the sequence. Each position in a read is associated with a quality score in ASCII format. It can be converted to a Phred score format by subtracting an offset value (33 in modern Illumina sequencers) from the numeric value of the ASCII character in the QUAL field. A higher Phred score in a read position corresponds to a lower probability of an incorrect base call in that position.

Two methods of consensus read collapse have been implemented in the pipeline:

- 1) The nucleotide with the highest frequency is taken as the consensus. If there are more than two types of nucleotides in one position (for example, A, C and G at the same coordinates), the position is removed to increase stringency and decrease the number of positions where mapping is ambiguous. In case of a tie in frequency, the position is also discarded by default. The mean quality score (Phred score) per position of the most frequent nucleotide is also calculated and the user can optionally also filter based on it.
- 2) The sum of the Phred scores of each nucleotide per position is calculated. The consensus nucleotide is the one with the highest sum of scores. The rest of the filtering steps are the same as the first method.

The second method was developed in case there is a position where one nucleotide exists in multiple low quality reads vs. another which exists in fewer high quality reads and we wanted the second. In practice, there is practically no difference between the two methods.

4. Transcriptional error detection

Positions covered by less than five consensus reads are discarded. Positions of known RNA editing sites are annotated through the REDportal database²⁸⁵ and removed. If a mismatch between the consensus reads has a frequency of more than one, then the position is discarded to avoid detection of errors caused by DNA polymorphisms or DNA damage. All other mismatches among the consensus reads are retained and their positions, type and surrounding nucleotides are stored. Error frequencies are detected by (a) dividing the number of genomic positions with mismatches in the consensus reads by the total number of positions that pass the filtering criteria and (b) dividing the number of mismatched nucleotides in the consensus reads by the total number of nucleotides. Method (a) was exclusively used for this analysis since the rates detected with (b) were highly correlated with the read coverage in each cell.

The calculation of the error rate per individual cell is noisy, because scRNAseq data are rife with dropout events and the low coverage increases the possibility that a transcriptional mistake will not be captured. If two samples differ in coverage, then the sample with a lower number of reads is more likely to have cells with zero transcription error rate. The mean or median error rate for a custom group of cells, either the whole sample or a cluster, can also be calculated from the output of the pipeline.

Localization of errors in genes and genome coverage plots

For genome annotation, the latest CellRanger GTF file was used. It is a subset of ENSEMBL annotations, with several gene biotypes removed (mostly small non-coding RNA). Genes were selected from that GTF using the R package *rtracklayer*²⁸⁶. To find the overlaps between the positions of transcription mismatches/consensus genome and the genome, we used the R package *GenomicRanges*²⁸⁷.

3.4 Discussion

scErrorRate is the first pipeline developed to measure the relative fidelity of RNA metabolism from scRNAseq data. It detected an increase in error rate with aging in mouse data and senescence in HUVEC cell culture. This is consistent with previous experiments that showed that transcription errors increase with age in yeast⁵. However, this is the first time that an increase in error rates with aging was demonstrated in mouse and human cells. Since problems in RNA metabolism can perturb several basic biological processes that are key to human aging, including proteostasis, metabolic pathways and NAD metabolism¹⁶², this finding can provide key insight into the biological basis of aging and age-related diseases.

scErrorRate also has the advantage that it operates on any UMI-based single cell RNA-seq dataset, unlike other approaches that require circle sequencing data. Additionally, the method can derive more accurate error rate estimates by summarizing information from a specific population of cells. When combined with cell clustering and cell type identification, it allows the comparison of error rates between different cell subpopulations.

One caveat is that the method cannot distinguish between true transcriptional misincorporations and errors caused by the process of reverse transcription during library construction. The reverse transcription process is error-prone, with the error rate dependent on the fidelity of the enzymes used. Because it lacks proofreading activity, a reverse transcriptase has an expected error rate of (6×10^{-5} to 6.7×10^{-4})²⁸⁸, which is much higher compared to the expected error rate of Pol II (4.9×10^{-6} to 5.6×10^{-5})^{162,175,179}. If the same RNA molecule is not transcribed multiple times, like in the rolling circle assays, then reverse transcription errors are expected to dominate in a standard RNA-seq library, which would make absolute estimation of error rate impossible. Nonetheless, when comparing samples from the same sequencing batch, the background noise caused by reverse transcriptase mistakes should be similar between samples. Relative differences in errors can be attributed to differential Pol II transcription accuracy and post-transcriptional RNA modifications. In any case, despite not correcting for RT errors, the error frequencies derived from polyA data are very similar to previously published data.

While the method was used specifically for aging data, its potential uses are broader. Transcription errors may contribute to diseases through multiple mechanisms. For instance, there is strong evidence that transcription mistakes have a direct, mechanistic effect on cancer²⁶⁶. They can facilitate oncogenesis by reducing the function of tumor-suppressor genes¹⁵⁹ or inducing cell cycle progression²⁸⁹. In later stages, they could potentially promote tumor evolution by increasing the phenotypic heterogeneity of tumor cells. Furthermore, cancers are characterized by increased levels of transcription and replication, which could make individual cells more susceptible to the consequences of error-prone transcription, since error-related backtracking can increase replicative stress through collisions of the different protein complexes²⁹⁰. Further study needs to be made on the contributions of transcriptional infidelity to cancer progression. By applying scErrorRate on the vast numbers of publicly available scRNAseq datasets from cancer tissue, we can tackle important questions in the field.

3.5 Contributions

The work described in this chapter is unpublished.

The group of Akis Papantonis generated the nascent scRNA-seq data. The provenance of the publicly available spleen data is detailed in the methods. I wrote this section and performed all the coding and analyses.

Chapter 4. General discussion

Analyzing the age-related changes to the molecular mechanisms of transcription and understanding their downstream effects on the composition of the transcriptome is an important avenue of research for both basic and clinical research. In the work presented here, we

1. observed an increase in elongation rate with aging and investigated its causes and consequences (Chapter 2)
2. created a pipeline to investigate RNA epimutations from single cell RNA sequencing data (Chapter 3)

4.1 The effects of aging-associated changes in transcriptional elongation on metazoan longevity

The estimated age-associated changes in Pol II speed are remarkably consistent across different metazoan species and cell types. Consequently, a mechanistic explanation of the changes would need to be a phenomenon that is observed in all eukaryotes. One of the potential causes that was investigated in this dissertation was the age-related alteration of the chromatin landscape. It is known that constitutive heterochromatin domains that are established early in development break down with aging and that senescent cells experience global heterochromatin loss.^{291,292} A global reduction in heterochromatin could increase the average speed of Pol II by relaxing the chromatin.

This is further supported by the fact that a global reduction in core histone proteins is one of the hallmarks of aging⁴. Replicative aging in budding yeast is accompanied by a dramatic reduction in histone protein levels²²². This reduction is also observed in aging worms²⁹³, senescent human cells²⁹⁴ and replicative aging human fibroblasts²⁹⁵. We showed that histone overexpression can increase chromatin compaction and decrease transcription speed, so it is reasonable to assume that the age-associated global decrease in histone protein levels would have the opposite result. We also showed that H3 overexpression increases proliferative lifespan in IMR90 cells, in agreement with previous results in worms²⁹⁶ and yeast²²².

It should be noted however that our work has not proved that histone loss or changes in chromatin density is the primary cause for the observed speed increase in all studied organisms. While the loss of core histones accompanies aging in many organisms, it still hasn't been observed *in vivo* in mitotic mammalian transcripts. MNase-seq in aged murine liver did not find any global reduction in nucleosome occupancy²⁹⁷. This result agrees with another recent study in multiple mouse tissues which actually showed that there was no drastic decrease in H3 expression levels²⁹⁸. Even if there is a global reduction in core histone expression with aging in mammals in some tissues, we haven't conclusively demonstrated that it would be sufficient to increase transcription speed.

4.1.1 Limitations

Some of the RNA Pol II-associated factors that control elongation rate also have an effect on aging and longevity. MYC, one of the four Yamanaka factors, is required for fast transcription elongation⁷⁹ and its depletion increases longevity and healthspan²⁹⁹. Depletion of the RecQ helicase RECQL5, which maintains genome stability by taking part in multiple DNA metabolic processes, increases elongation speed⁷⁵. PNUITS-PP1, which negatively regulates elongation speed, is repressed in senescent cells³⁰⁰. The change in expression of these proteins during aging would potentially directly affect Pol II speed, explaining at least partially some of the changes that we observe. Additionally, given that these factors are associated with various crucial cellular processes, some of the observed effects of the acceleration of transcriptional elongation might be caused by the age-related changes in their expression.

The slope used to estimate Pol II speed is also dependent on the initiation rate n , defined as the number of polymerases initiating transcription per unit of time. Slope and n are directly correlated, so if n significantly decreases with aging, we would also expect to see on average shallower slopes of read distributions. Reduction of n would mean that either:

- 1) the loading of the polymerase and the assembly of the pre-initiation complex happens more slowly
- 2) promoter-proximal pausing lasts longer, presumably because of less efficient pause release.
- 3) promoter-proximal termination occurs more often, decreasing the number of elongating molecules per unit of time
- 4) some combination of the three

According to a recent study by Bozukova et al³⁰¹, a strong decrease in promoter-proximal pausing was observed in NET-SEQ data from aged murine liver. Even though initiation rates cannot be directly estimated, the data points towards increased initiation rates and reduced proximal pausing in aged tissues, which would globally decrease estimated elongation rates. However, the same data also suggests that the frequency of promoter-proximal termination increases with aging. Further work would need to be done to understand the impact of initiation on our results.

4.1.2 Future directions

The focus of our research so far was on global increase of elongation rate, since our method is too noisy to accurately estimate speed changes on the gene level. However, chromatin and histone modifications can change locally, depending on the cellular environment. This can affect elongation rate and thus the production of specific transcripts, by altering splicing, termination or circular RNA formation. Understanding the mechanisms behind such changes and the stimuli that trigger them would provide insight into the effects of regulated speed change. It would also be interesting to investigate global and local variability in Pol II speed during development, given the evidence that altered elongation rate can be embryonically lethal¹³¹ in mice and non-viable in plants¹³².

Furthermore, it would be interesting to investigate how epigenetic factors affect Pol II speed using our total RNA-seq dataset. There are publicly available databases containing the locations of epigenetic modifications and ATAC-seq peaks, such as the one provided by the ENCODE project³⁰², that can be used for this purpose. Similarly, we can investigate how the presence or absence of DNA binding proteins affects the estimated elongation rates.

4.2 Age-related changes in transcriptional fidelity across tissues

In chapter 3, we presented ScErrorRate, the first method to estimate transcriptional error rates from single-cell RNA-seq data. By taking advantage of the presence of unique molecular identifiers in the dataset, our method avoids the laborious process of tandem repeat methods and is applicable to any public or newly generated UMI-based scRNAseq dataset.

Despite the expected increase in noise by not correcting for reverse transcriptase errors, the error spectrum we derive from polyA data is similar to previous results that use circle sequencing to account for the infidelity of reverse transcription. This is surprising, given that the estimated error rate of RT is at least an order of magnitude greater than Pol II. It is worth noting however that a fully satisfactory method to determine fidelity of RNA-dependent DNA synthesis by RTs is still missing. Most fidelity estimates either lack a way to differentiate between transcription and reverse transcription errors^{303–305} or have been obtained from DNA-dependent DNA polymerization assays of the enzyme^{288,306}. A recent study based on M13 lacZ forward mutation assays showed that RNA-dependent DNA polymerization error rates had values in the range of 2.5×10^{-5} to 3.5×10^{-5} . This range is closer to the reported RNA polymerase transcription error rates, which indicates that the errors introduced by RNA polymerase, either T7 or the equivalent enzymes used for *in vitro* transcription, inflate the estimated inaccuracy of reverse transcription.

The difference we observe in the error spectrum of nascent RNA-seq data is also remarkable. It cannot be attributed solely to differences in RNA editing. Firstly, from what is known, RNA editing takes place mostly co-transcriptionally^{8,307} and is responsible for changes in alternative splicing^{308,309}. Second, while A-to-I editing events are prevalent, there is much less evidence of widespread C-to-U editing, which is the most common type of error in the spleen data and in previous publications. Further studies are necessary, altering the expression of deaminase enzymes or modulating nitrosative stress to clarify how much deamination contributes to C-to-U RNA mutations.

This is also the first time increased genome-wide transcriptional infidelity with aging and senescence has been directly measured in higher organisms. Previous research has shown that an increase in transcriptional mistakes increases cytotoxic stress and negatively affects lifespan in yeast⁵. If this is also the case for humans, it could have significant implications for aging-related diseases. There are several potential mechanisms that could explain why aging leads to decreased transcriptional fidelity. One of them is oxidative stress. DNA damage accumulates in cells with aging³¹⁰, partially because age-related defects in respiration create oxidative DNA lesions^{311,312}. These lesions can significantly increase misincorporations during transcription^{13,313}. For example, the modified guanine 8-oxoguanine, which results from oxidation of normal guanine, causes a C-to-A transversion when transcribed³¹⁴. This transversion has been associated with a reduction in splicing fidelity³¹⁵ and α -synuclein aggregation³¹⁶. Another potential mechanism is the age-related increase of Pol II speed; increased speed of elongation is linked with reduced Pol II precision¹⁴¹.

4.2.1 Limitations

ScErrorRate, while functional, is not yet well optimized. Running it on a single sample currently takes more than a day. The most time consuming step is the pileup process, which is currently processed in R. Implementing the step in C would drastically reduce processing time and make the pipeline more accessible for broader use.

Regarding the effects of aging on transcriptional fidelity, our results are still preliminary. We have only tested scErrorRate on two samples, so there might be unknown technical issues that influence our results. Further testing on the multiple available aging datasets is required to strengthen our conclusions. The rest of the Tabula Muris Senis dataset would be a good candidate; it provides a deep characterization of the aging transcriptome on a single cell level, including animals of multiple ages, from 1 months to 30 months old. The Calico study¹⁹ is also of interest, as it contains kidney, lung and spleen 10x sequencing data from young and old mice.

4.2.2 Future directions

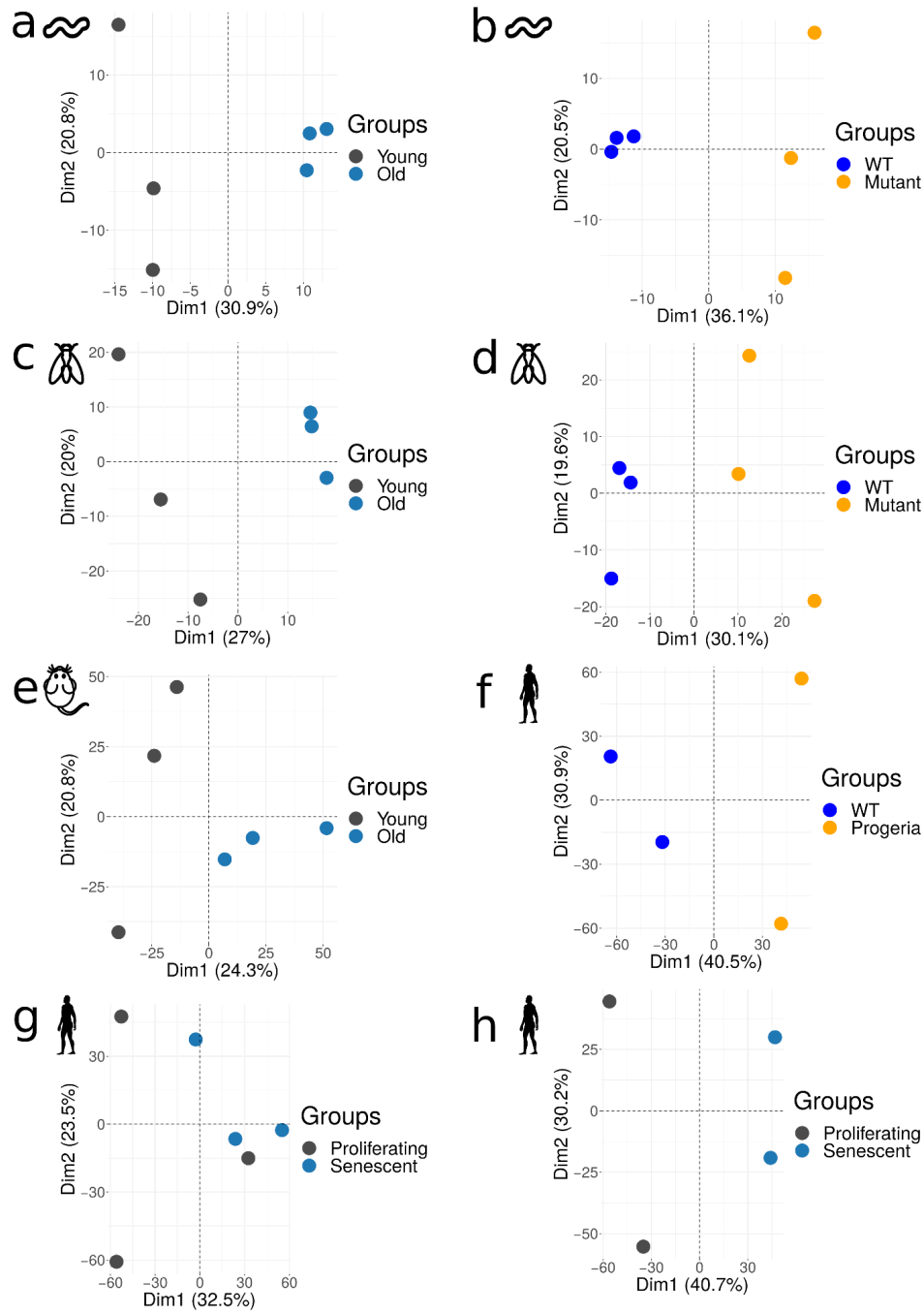
Potential integration of other information layers would provide more insight into the mechanisms behind transcription errors. Recently, single cell ATAC-seq has been applied in droplet-based platforms³¹⁷, allowing the integration of expression data with an assessment of the chromatin status of each cell. Applying our method on combined scRNAseq/single-cell ATAC-seq data would elucidate the effects of DNA accessibility on the error rate of transcription.

There is some evidence¹⁶⁰ that transcriptional infidelity increases splicing defects. 10x sequencing is not suitable for investigating the effects of Pol II errors on splicing, as it provides limited coverage of the transcripts. New methods have been developed in the last few years that combine full-length transcript coverage and UMI information, like Smart-seq3¹⁸⁷. Using scErrorRate on reads spanning exon-intron junctions in combination with an isoform-level analysis could expand our understanding of the splicing mechanisms.

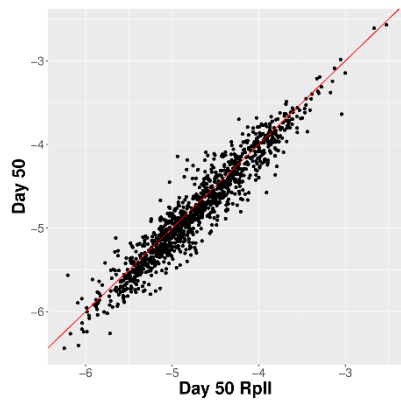
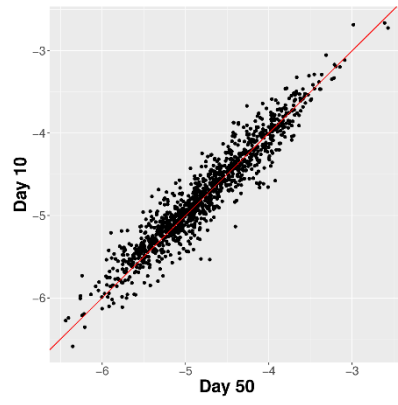
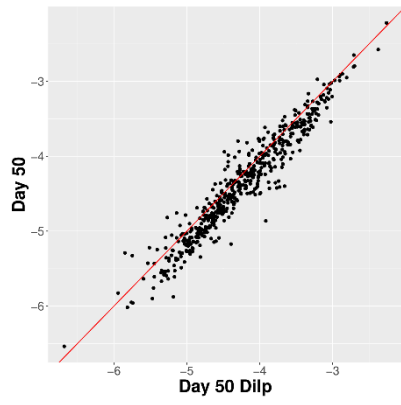
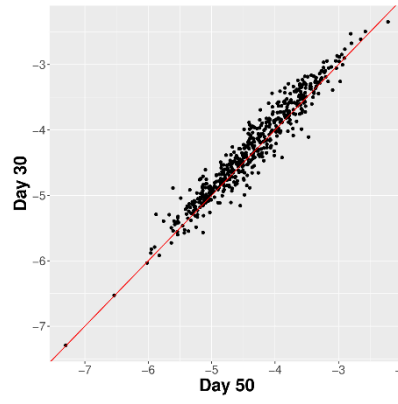
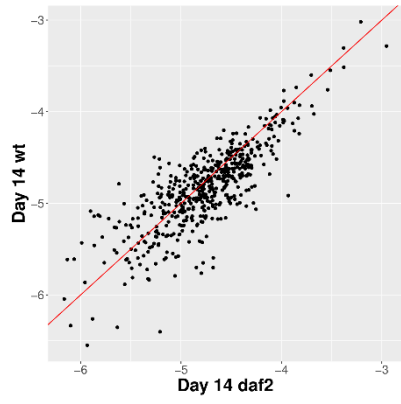
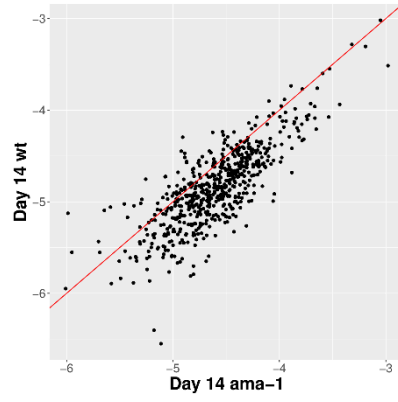
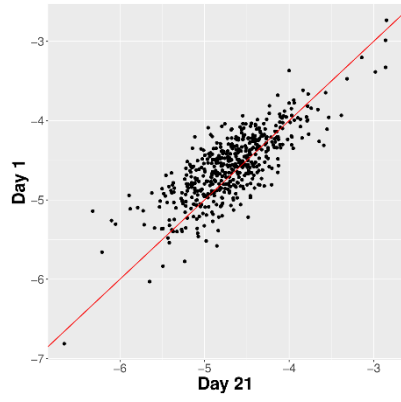
Changes in elongation rate could potentially have an effect on transcriptional fidelity, but this has never been experimentally demonstrated *in vivo* in a conclusive fashion. If we perform single-cell analysis on Pol II slow or fast mutants, not only would we obtain insight on a single-cell level on the molecular mechanisms behind the changes we observe in lifespan, but we could also use scErrorRate to see how alterations in elongation rate affect transcriptional errors.

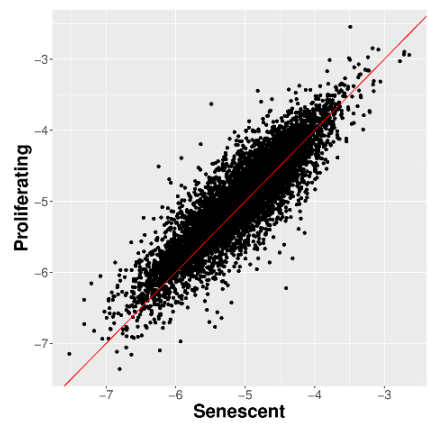
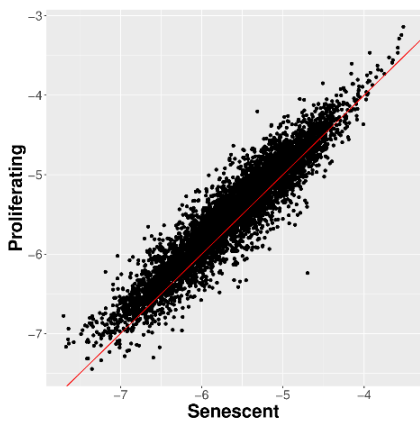
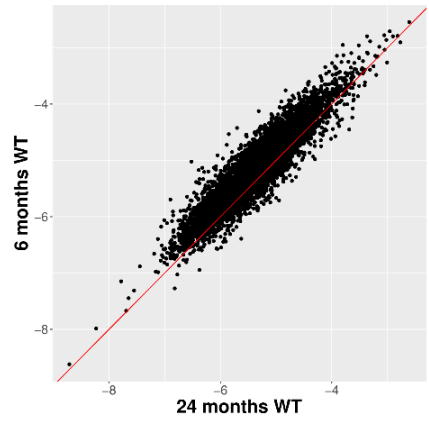
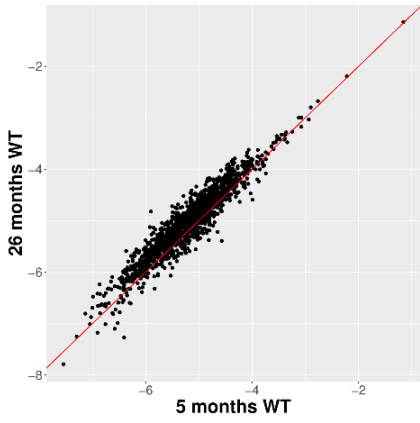
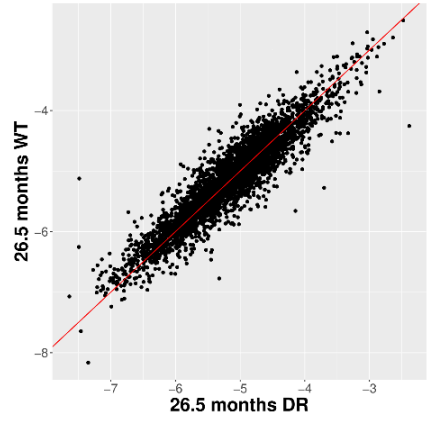
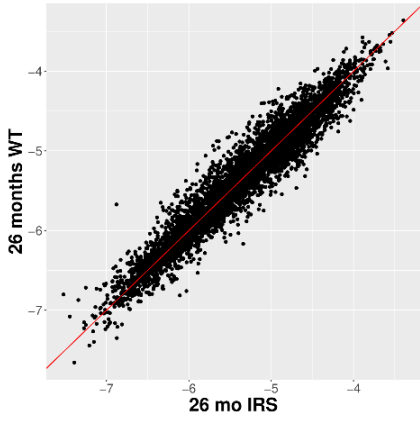
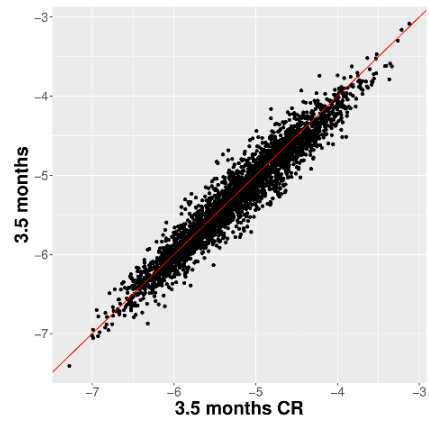
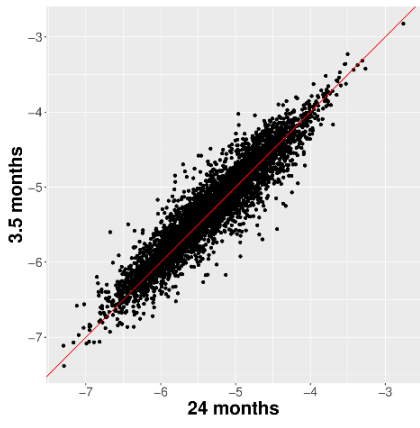
Finally, there is no reason to limit the application of our method to aging and aging-related diseases. It can be used to map transcription errors in different species and tissues, under different conditions (environmental stresses or DNA mutations) and to investigate the consequences of specific errors. It would be interesting to discover the phenotypic effects of transcriptional infidelity and try to uncover the mechanisms underlying them.

Appendix A: Supplementary Figures

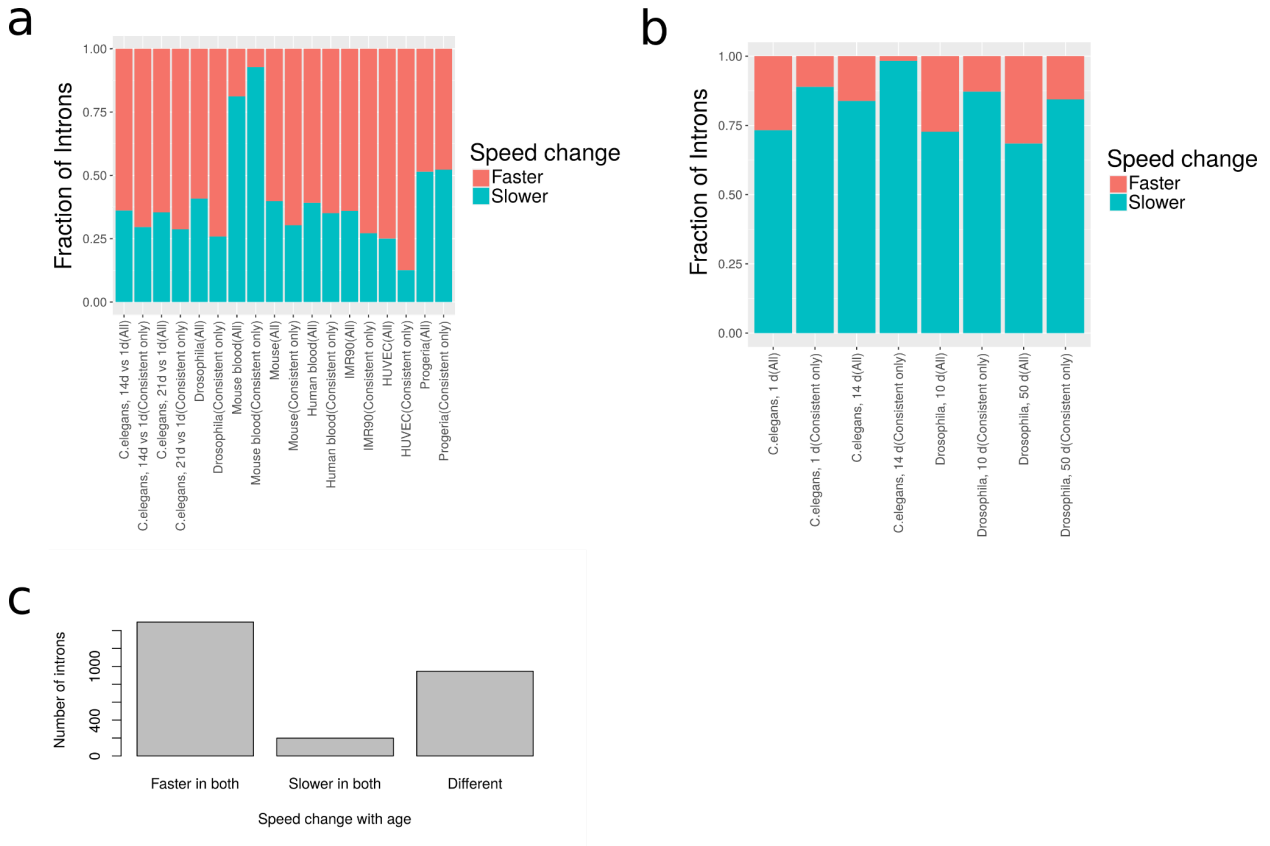


Supplementary Figure 2.1: PCAs of slopes of intronic read distribution. Principal component analysis (PCA) of the slopes of *C. elegans* ((a) wt 21 d vs 1 d; (b) 14 *ama-1(m322)* d vs wt 14 d), *D. melanogaster* ((c) wt heads 50 d vs 10 d, (d) Rpl1215⁴ heads 50 d vs wt 50 d), *M. musculus* ((e) kidney: 24 mo vs 3 mo), *H. sapiens* ((f) Progeria: WT vs progeria mutants, (g) HUVEC and (h) IMR90: Senescent vs Proliferating).

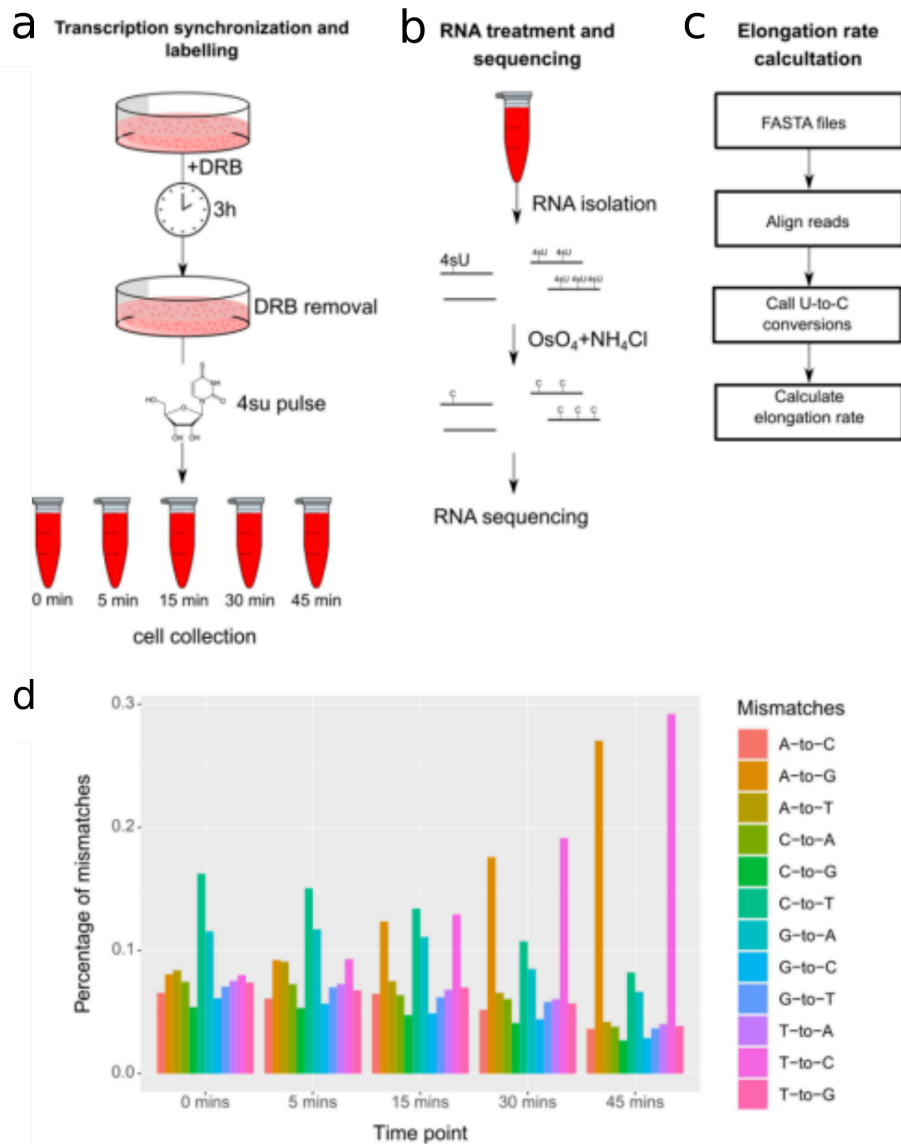




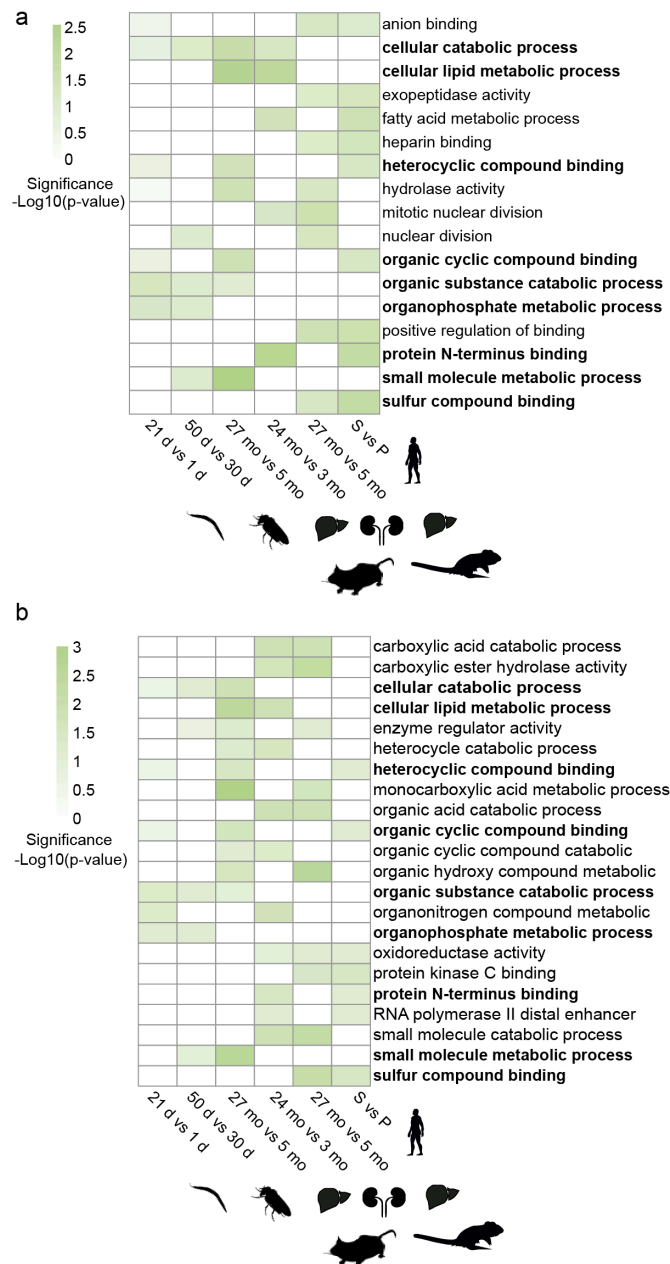
Supplementary Figure 2.2: Scatterplots of intronic slope ($-\log_{10}$) for each condition and species (*C. elegans*, *D. melanogaster*, *M. musculus*, *R. norvegicus*, *H. sapiens*).



Supplementary Figure 2.3: Consistency of RNA Pol-II speed changes. (a) Change of elongation rate with aging or senescence in introns of *C. elegans*, *D. melanogaster*, *M. musculus* and *H. sapiens*, before and after filtering for introns that consistently change in speed in all replicates. (b) Change of elongation rate with mutations that slow down the speed of RNA-Pol-II in introns of *C. elegans* and *D. melanogaster* before and after filtering for introns that consistently change in speed in all replicates. (c) Comparison of the change of elongation rate with aging between IMR90 and HUVEC using the same introns.

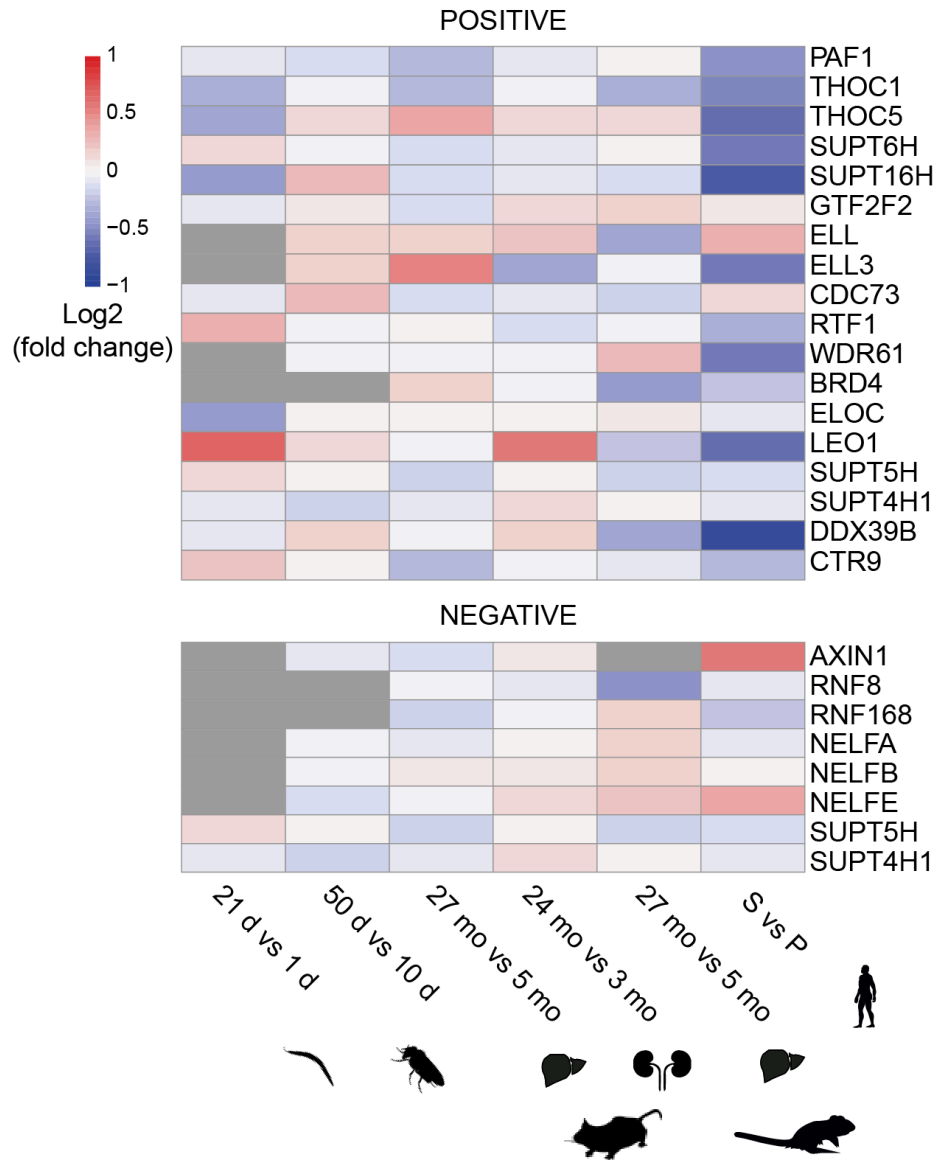


Supplementary Figure 2.4: 4SU-DRB labeling and TUC conversion to calculate RNA-Pol-II elongation rate. (a-c) Schematic representation of the 4SU-DRB labeling (a), TUC conversion (b) and elongation rate calculation (c). **(d)** Percentage of mismatches in every time point of the experiment (0 mins, 15 mins, 30 mins, 45 mins) in one of the proliferating replicates. There is a noticeable increase in A-to-G and T-to-C mismatches in the last two time points.

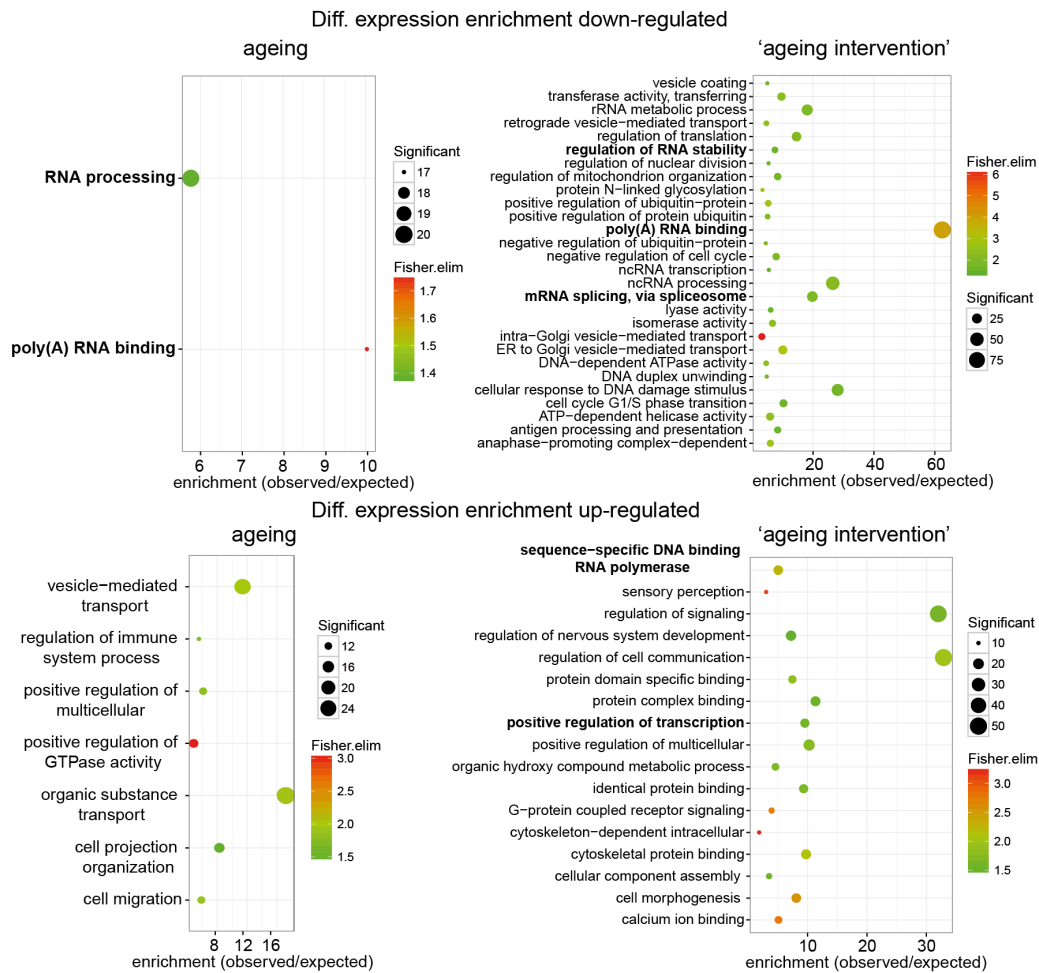


Supplementary Figure 2.5: Genes with increase in Pol-II speed are associated with metabolism and catabolism related pathways. GO enrichment analysis of genes with increased Pol-II speed across species: *C. elegans* (21 d vs 1 d), *D. melanogaster* (heads: 50 d vs 30 d), *M. musculus* (kidney: 24 mo vs 3 mo), *R. norvegicus* (liver: 24 mo vs 6 mo), *H. sapiens* (IMR90: Senescent vs Proliferating). GO enrichment of (a), top 200 (b), top 300 genes with an increase in Pol-II speed change for each species (common terms between the two sets in bold). Color scale indicates the significance of the enrichment (all GO terms enriched with p-values below 0.05, with at least 10 significant genes for each GO categories, Fisher elim test).

REGULATION OF DNA TEMPLATED TRANSCRIPTION ELONGATION

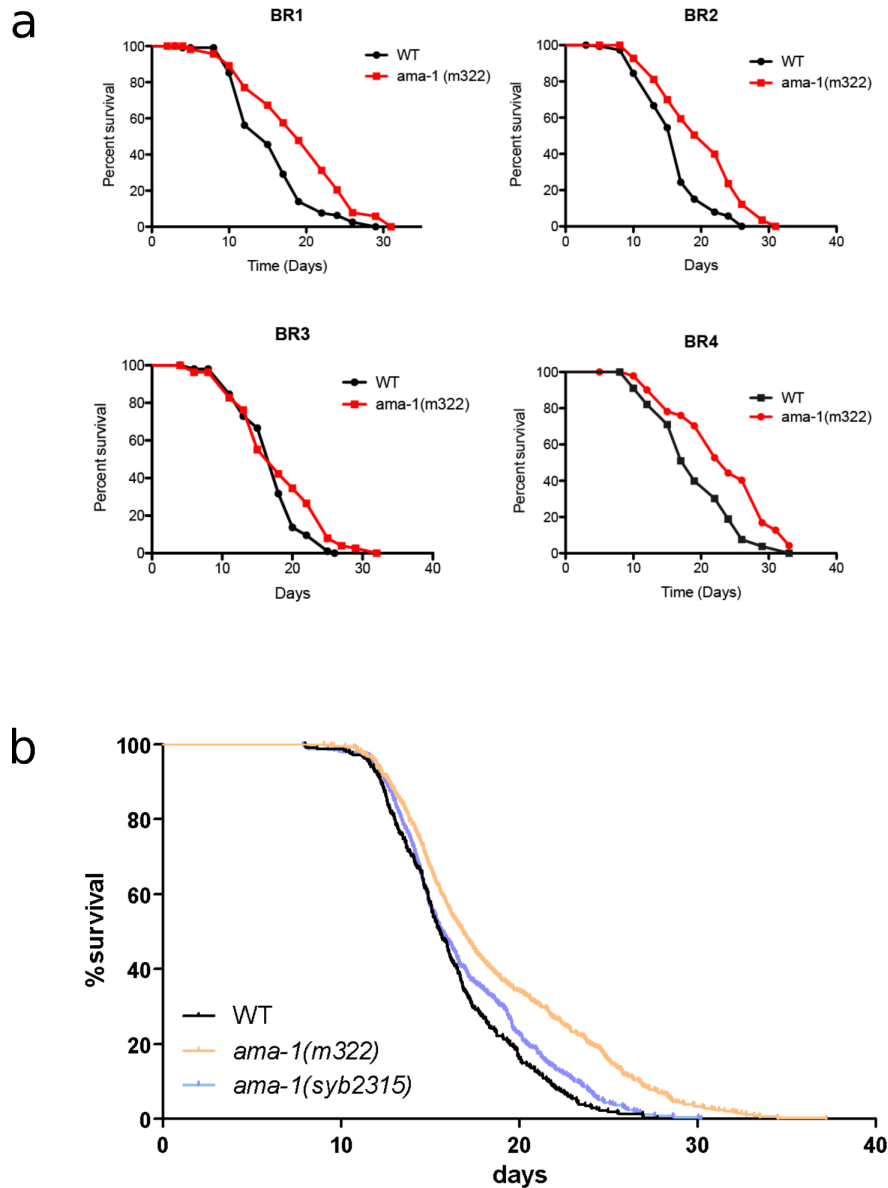


Supplementary Figure 2.6: Heatmap of differential expression (\log_2 fold change) of MSigDB (61) annotated genes for 'regulation of DNA templated transcriptional elongation'. **Top: activators of transcriptional elongation (POSITIVE); **Bottom:** repressors of transcriptional elongation (NEGATIVE). Data shown for *WT* aging time courses: worm (21 d vs 1 d), fly heads (50 d vs 10 d), mouse liver (27 mo vs 5 mo), mouse kidneys (24 mo vs 3 mo) and human fibroblast cell line (IMR90: Senescent vs proliferating).**

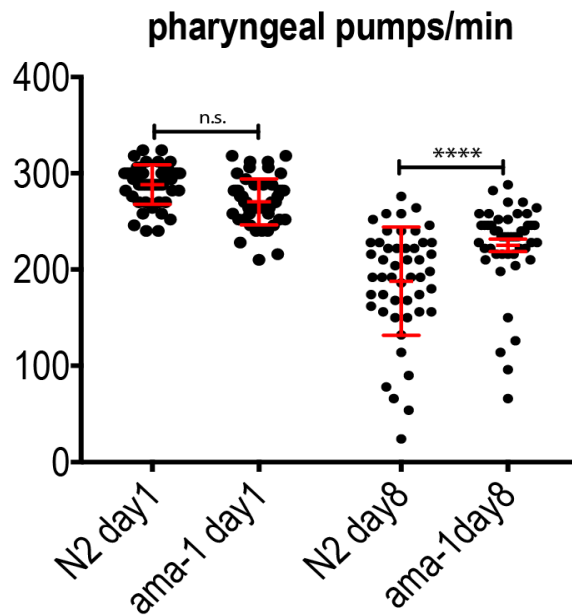


Supplementary Figure 2.7: Functional enrichment for across-species differential expression analysis.

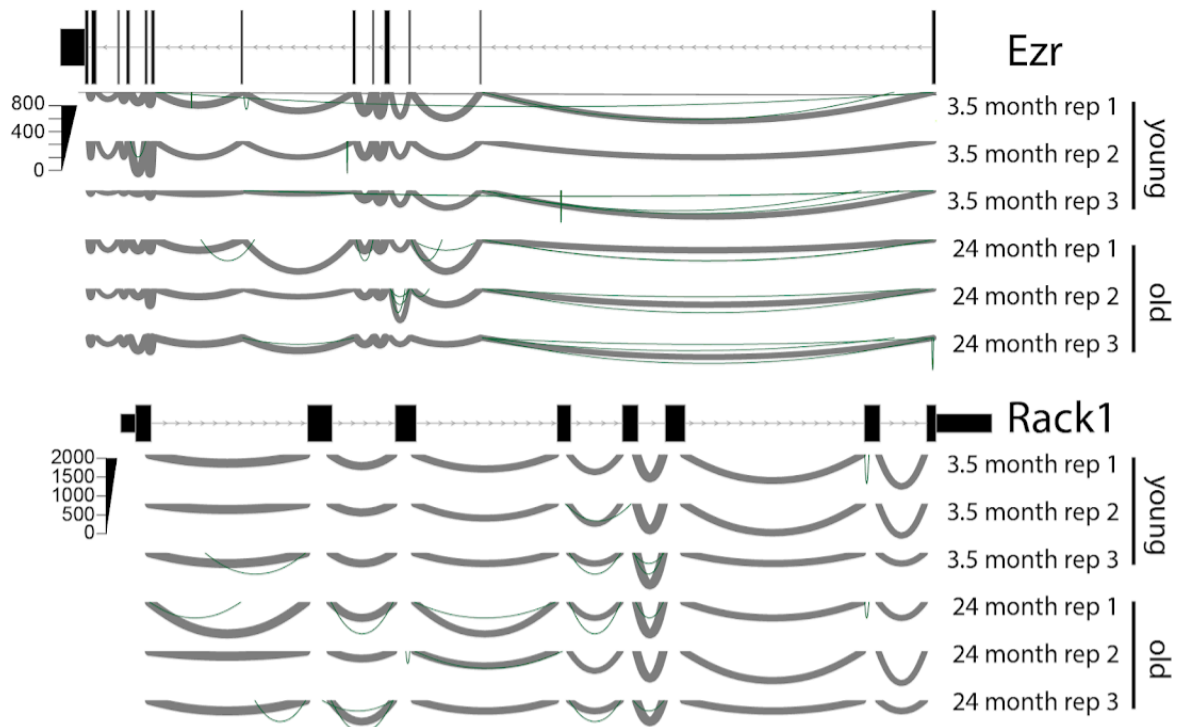
GO enrichment for consistently down-regulated (top) or up-regulated (bottom) genes across species during aging (left) or ‘aging intervention’ (right) (aging up-regulated: 92 genes; aging down-regulated: 71 genes; ‘aging intervention’ up-regulated: 164 genes; ‘aging intervention’ down-regulated: 473 genes; as background for the enrichment analysis a set of 4784 orthologue genes between *H. sapiens*, *R. norvegicus*, *M. musculus*, *D. melanogaster*, *C. elegans* was used. All p-values *P < 0.05, significant genes > 10, fisher elim test). GO terms related to transcription and splicing are indicated in bold.



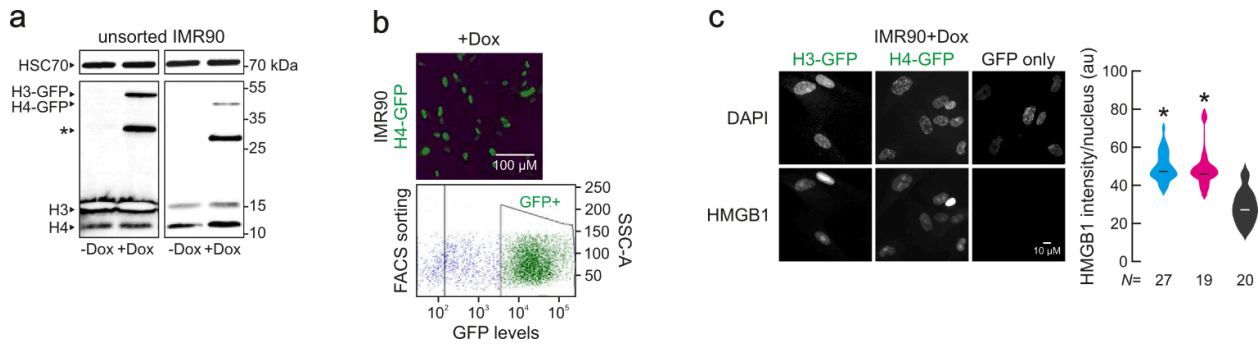
Supplementary Figure 2.8: Slowing down Pol-II in *C. elegans* increases lifespan. (a) Survival of wild-type and *ama-1(m322)* mutant worms conferring a slow Pol-II elongation rate (4 replicates, BR1:1.267, $P < 0,0001$; BR2:1.23, $P < 0.0001$; BR3:1, $P = 0.0342$; BR4:1.263, $P < 0.0001$, log-rank test, Mantel-cox). **(b)** *C. elegans* lifespan analysis after CRISPR/Cas9 mediated reversion of the slow RNAPII mutation. Survival curves of the strain harboring the slow RNAPII mutation (*ama-1 m322*) and wild-type controls compared to worms after CRISPR/Cas9 engineered reversion of the slow mutation back to the wild type allele (*ama-1 syb2315*). Animals with slow Pol-II have a significantly increased lifespan. CRISPR/Cas9 engineered reversion restored lifespan essentially back to wild-type levels. (3 replicates; $n > 300$ per strain).



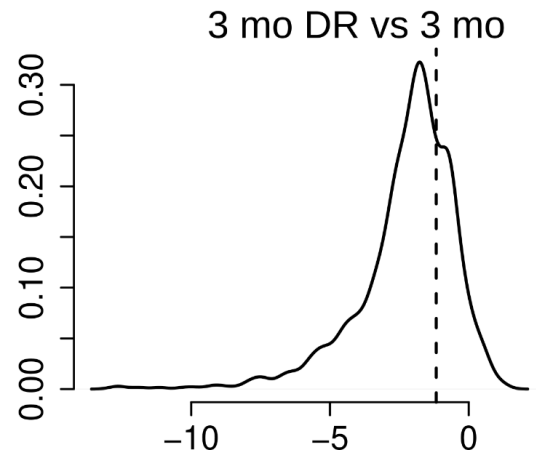
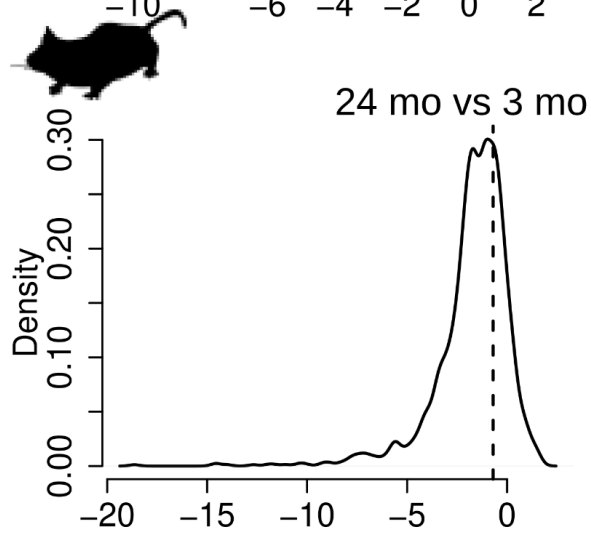
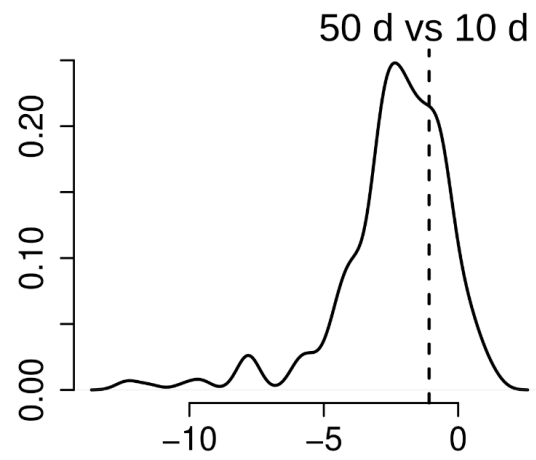
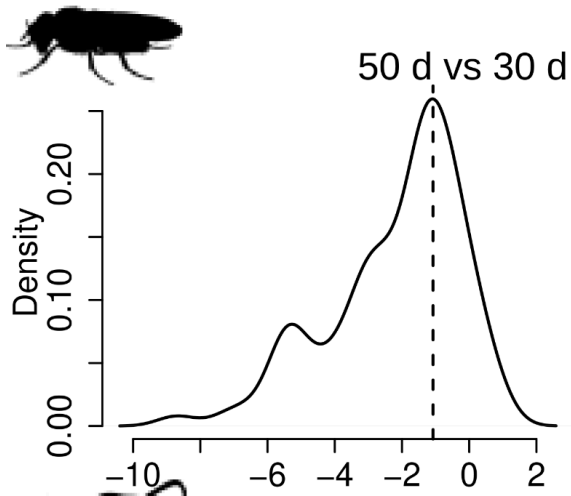
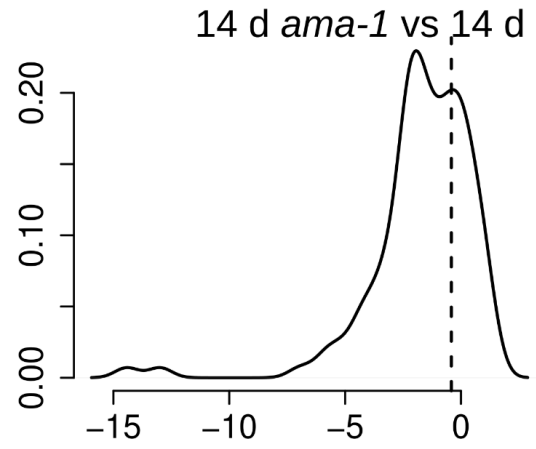
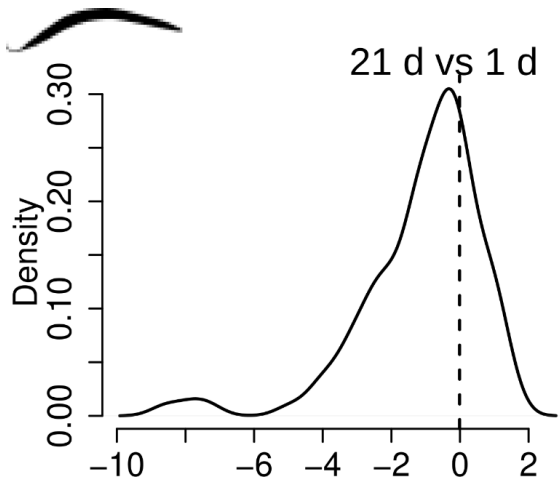
Supplementary Figure 2.9: Slowing down Pol-II in *C. elegans* ameliorates the age-related decline in pharyngeal pumping rates. Pumping rates of wild type N2 and *ama-1* mutant worms were measured on day 1 and day 8. Pumping rates were not significantly different on day 1, but *ama-1* worms showed higher pumping rates compared to wild types on day 8, suggesting that the mutant worms are healthier at old age.

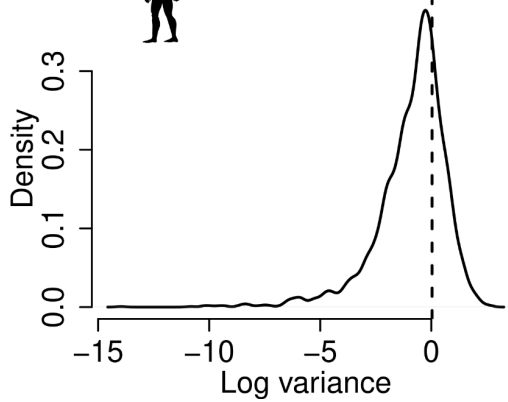
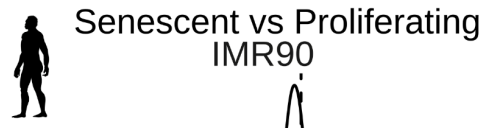
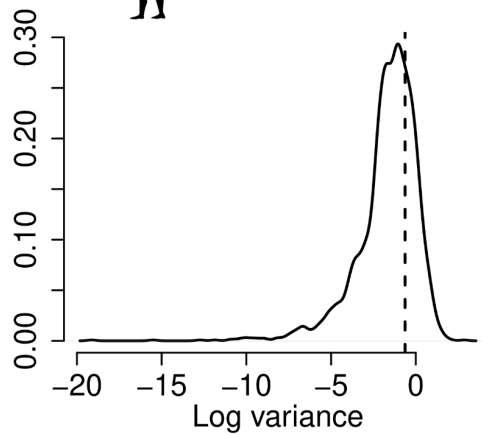
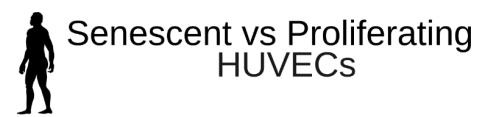
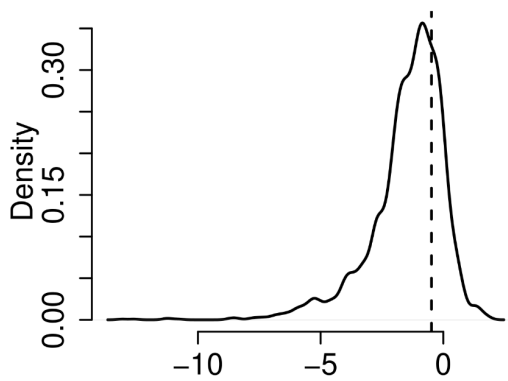


Supplementary Figure 2.10: Examples of rare splice site changes for gene *Ezr* and *Rack1* with 3 replicates young (3.5 month) and old (26 month). Line thickness encodes the number of reads supporting this junction. Rare splice sites are shown in green.

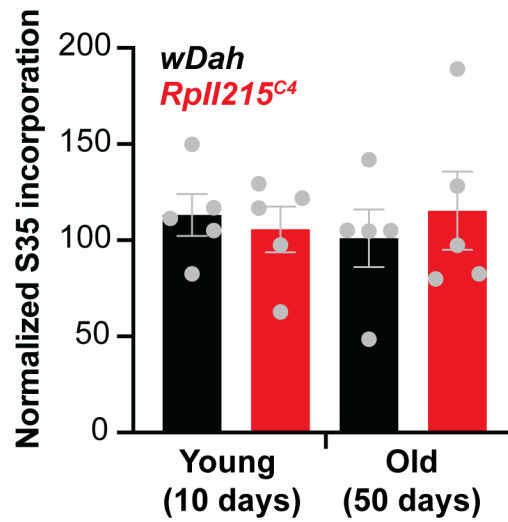


Supplementary Figure 2.11: H3-GFP and H4-GFP overexpression in IMR90 cells. (a) Western blot experiments confirm the overexpression of the H3-GFP and H4-GFP proteins. **(b)** Visual confirmation of the Dox induction of H3/H4 expression and FACS sorting of GFP-positive cells. **(c)** Typical immunofluorescence images of H3-GFP, H4-GFP and control IMR90 cells (*left*) show increased DAPI levels in histone overexpression nuclei. Violin plots (*right*) quantify this reduction. N specifies the number of cells analyzed per condition.



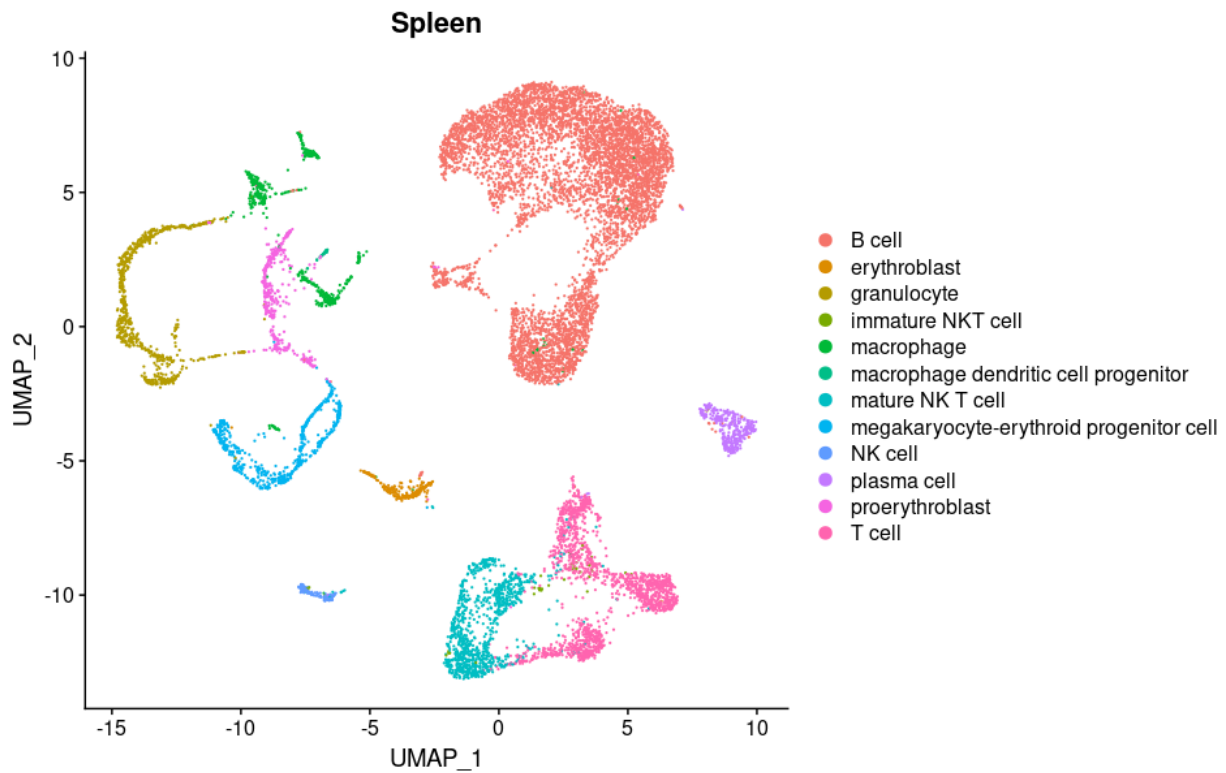


Supplementary Figure 2.12: Variation of Pol-II elongation speed changes for different introns of the same gene. Distribution of variances of Pol-II speed estimates (slope per intron) for introns within the same gene. Average variance of speed estimates across all introns (i.e. between genes; global average) is shown as a dashed vertical line for *C. elegans* (21 d vs 1 d; 14 *daf-2* d vs 14 d), *D. melanogaster* (heads 50 d vs 30 d; 50 d vs 10 d), *M. musculus* (kidney: 24 mo vs 3 mo; 3 DR mo vs 3 mo), *R. norvegicus* (liver: 24 mo vs 6 mo), *H. sapiens* (Umbilical vein endothelial (HUVECs); fibroblast fetal lung (IMR90): Senescent vs Proliferating). The vast majority of intra-gene variances are below the average inter-gene variance, suggesting that introns of the same gene have coupled Pol-II elongation speeds.

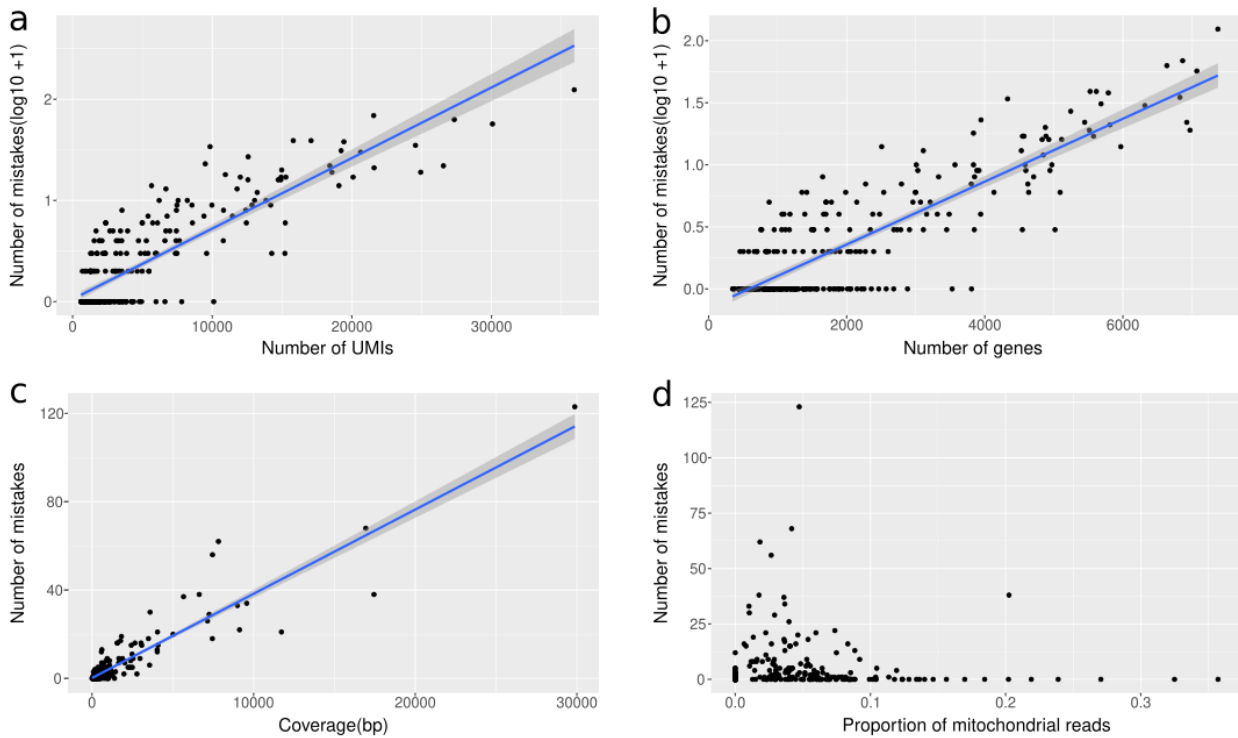


Supplementary Figure 2.13: Protein biosynthesis rates do not change with aging in *Drosophila*.

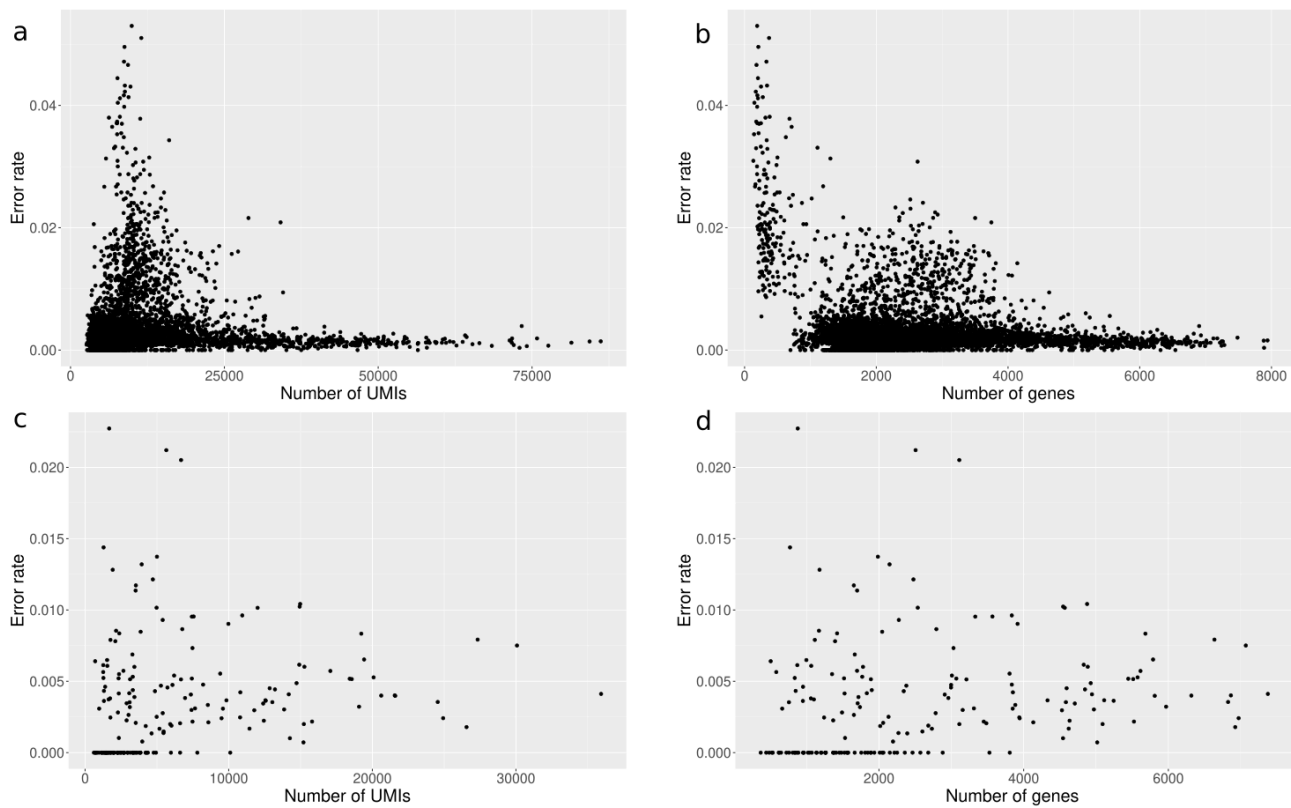
Ex-vivo S35 incorporation assay shows no significant difference in translation rates in female fly heads between wDah control and Rpl215C4 mutants both at young (10days) and old age (50 days). N=5 biological replicates with 25 heads per replicate.



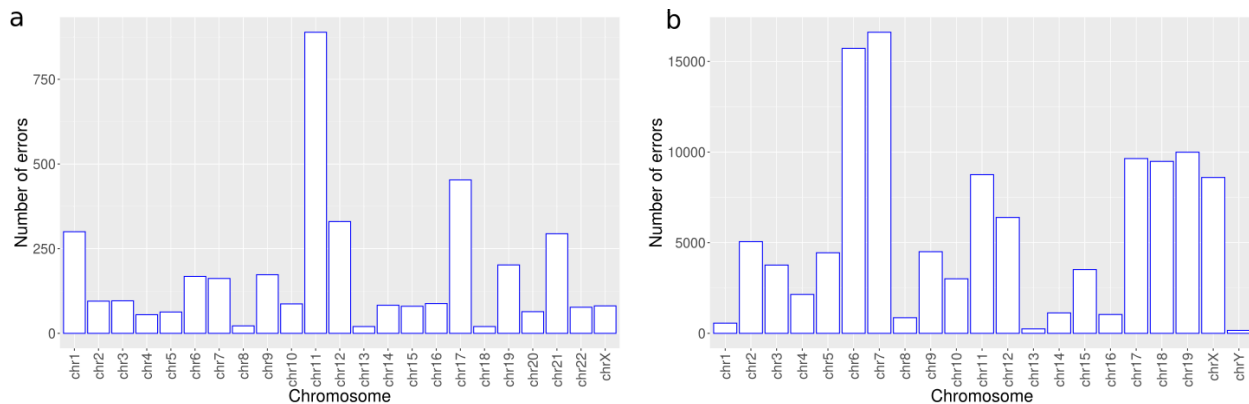
Supplementary Figure 3.1: Visualization of the Tabula Muris Senis spleen dataset. Despite being realigned with STARSolo, cells keep their original clustering.



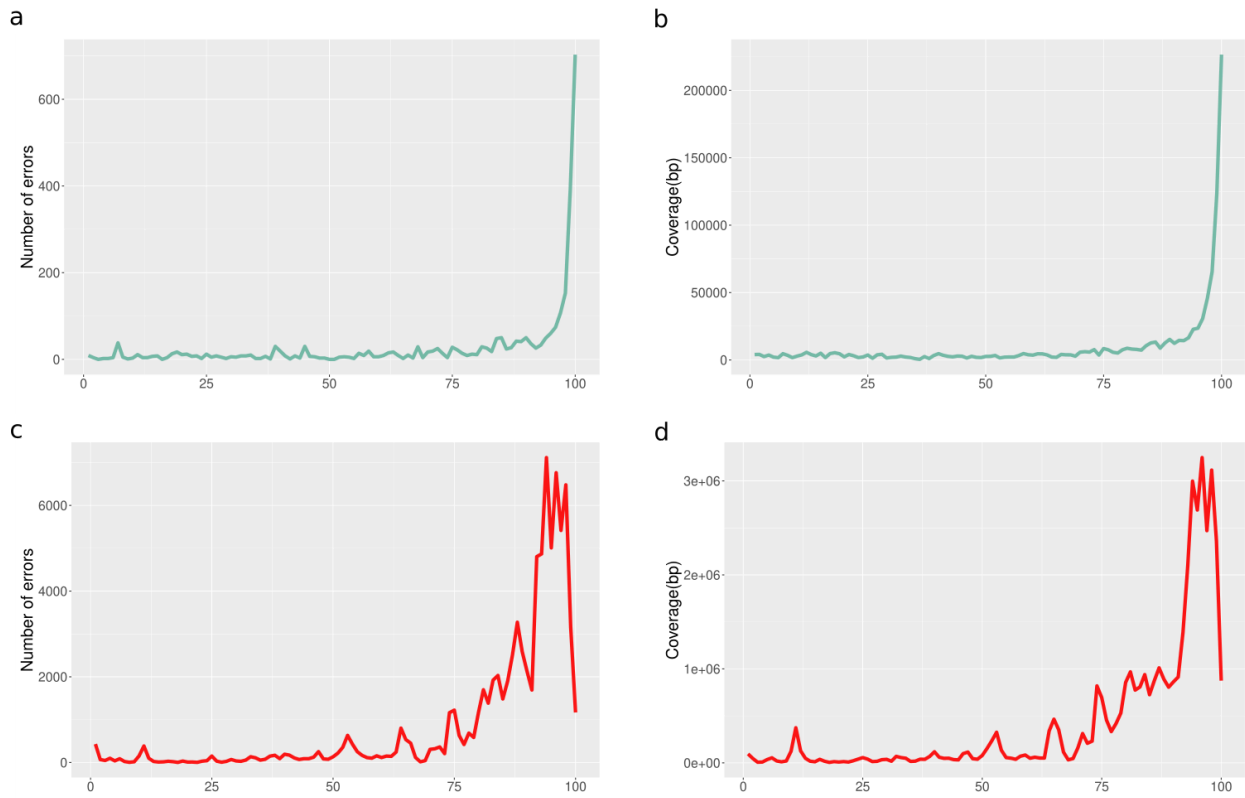
Supplementary Figure 3.2: Association between numbers of transcription errors per cell and cellular features. The number of detected mistakes per cell in the HUVEC dataset is highly correlated with (a) the number of UMIs per cell (Pearson correlation = 0.812, p-value < $2.2e^{-16}$), (b) the number of expressed genes per cell (Pearson correlation = 0.717, p-value < $2.2e^{-16}$) and (c) the coverage of consensus sequences per cell (Pearson correlation = 0.920, p-value < $2.2e^{-16}$). It is uncorrelated (Pearson correlation = 0.0029, p-value = 0.962) with mitochondrial expression (d).



Supplementary Figure 3.3: Association between error rate per cell and cellular features. The error rate per cell in the spleen dataset is very lowly correlated with (a) the number of UMIs per cell (Pearson correlation = 0.0856, p-value < $2.2e^{-16}$) and (b) the number of expressed genes per cell (Pearson correlation = -0.1422, p-value < $2.2e^{-16}$). Similar low correlations with the number of UMIs (c, Pearson correlation = 0.165, p-value = 0.025) and genes per cell (d, Pearson correlation = 0.173, p-value = 0.019) are found in the HUVEC dataset.



Supplementary Figure 3.4: Count of transcription errors across human (a) and mouse (b) chromosomes.



Supplementary Figure 3.5: Gene coverage of transcriptional errors. The coding genes were binned in 100 bins from 5' end to 3' end and the number of detected errors (a, c) and consensus sequence positions (b, d) from every HUVEC and spleen cell were counted.

Appendix B: Supplementary Tables

Species	Comparison	Distance from promoter	Intron length	Gene expression log2FC	Circular RNA index
<i>C. elegans</i>	21 d vs 1 d	0.014	0.071	-0.266	-0.085
<i>C. elegans</i>	14 d ama-1 vs 14 d wt	0.008	-0.110	-0.247	0.160
<i>C. elegans</i>	14 d daf2 vs 14 d wt	0.020	0.007	-0.278	0.130
<i>D. melanogaster</i>	50 d vs 10 d	-0.036	0.019	-0.023	0.019
<i>D. melanogaster</i>	50 d RplI215 vs 50 d	0.002	-0.091	-0.161	-0.043
<i>D. melanogaster</i>	50 d dilp 2,3-5 vs 50 d	-0.133	-0.170	0.041	-0.092
<i>M. musculus</i>	24 mo vs 3.5 mo	0.011	0.043	-0.230	-0.045
<i>H. sapiens</i>	Senescent vs Proliferating (IMR90)	-0.021	0.046	-0.274	0.184
<i>H. sapiens</i>	Senescent vs Proliferating (HUVEC)	0.014	0.036	-0.289	0.020

Supplementary Table 2.1: Table of correlations (Pearson correlation) between the change in elongation rate and characteristics of the introns in which the elongation rate was measured in selected RNA-SEQ datasets.

Species	Tissue	Enrichment protocol	Sequencing parameters Paired-/single-end, read length, millions of reads (M)	Time points ¹ and conditions
<i>C. elegans</i>	Whole body	TruSeq Stranded Total RNA Library	Paired-end, 75 bp, 25 M	Day 1 (WT), day 7 (WT), day 14 (WT), day 21 (WT)
	Whole body	TruSeq Stranded Total RNA Library	Paired-end, 75 bp, 25 M	Day 14 (WT, <i>daf-2(e1370)</i> , <i>ama-1(m322)</i>)
	Whole Body	TruSeq Stranded Total RNA library	Paired-end, 75 bp, 25 M	Day 1 (WT, <i>ama-1(m322)</i>)

¹ Triplicate except where mentioned otherwise.

<i>D. melanogaster</i>	Head	TruSeq Stranded Total RNA Library	Single-end, 100 bp, 37.5 M	Day 30 (<i>WT</i> , <i>dilp2,3-5</i>), Day 50 (<i>WT</i> , <i>dilp2,3-5</i>)
	Head	TruSeq Stranded Total RNA Library	Paired-end, 75 bp, 30 M	Day 10 (<i>WT</i> , <i>Rpl1215^{C4}</i>), day 50 (<i>WT</i> , <i>Rpl1215^{C4}</i>)
<i>M. musculus</i>	Kidney	TruSeq Stranded Total RNA Library	Paired-end, 75 bp, 70 M	Month 3 (<i>WT</i>), month 24 (<i>WT</i>)
	Kidney	TruSeq Stranded Total RNA Library	Paired-end, 75 bp, 30 M	Month 3 (<i>WT</i> , <i>DR</i>) (4 replicates)
	Liver	TruSeq Stranded Total RNA Library	Paired-end, 75 bp, 37.5 M	Month 5 (<i>WT</i> , <i>DR</i>)
	Liver	TruSeq Stranded Total RNA Library	Paired-end, 75 bp, 37.5 M	Month 16 (<i>WT</i> , <i>DR</i>)
	Liver	TruSeq Stranded Total RNA Library	Paired-end, 75 bp, 37.5 M	Month 27 (<i>WT</i> , <i>DR</i>)
	Blood	TruSeq Stranded Total RNA Library	Paired-end, 75 bp, 70 M	Month 5 (<i>WT</i>), month 27 (<i>WT</i>)
	Hypothalamus	TruSeq Stranded Total RNA Library	Single-end, 100 bp, 30 M	Month 26 (<i>WT</i> , <i>IRS1^{-/-}</i>)
<i>R. norvegicus</i> (17)	Liver	TruSeq Stranded Total RNA Library	Single-end, 50 bp, 60 M	Month 6 (<i>WT</i>), month 24 (<i>WT</i>) (2 replicates)

	Brain ²	TruSeq Stranded Total RNA Library	Single-end, 50 bp, 20 M	Month 6 (<i>WT</i>), month 24 (<i>WT</i>) (6 replicates)
<i>H. sapiens</i>	Fetal lungs (IMR90)	Nascent RNA	Paired-end, 75 bp, 25 M	Early passage, late passage (2 replicates)
	Fetal lungs (IMR90)	TruSeq Stranded Total RNA Library	Paired-end, 75 bp, 50 M	Early passage, late passage (2 replicates)
	Umbilical vein endothelial (HUVECs)	Nascent RNA	Paired-end, 75 bp, 50 M	Early passage, late passage (2 replicates)
	Umbilical vein endothelial (HUVECs)	TruSeq Stranded Total RNA Library	Paired-end, 75 bp, 100 M	Early passage, late passage
	Fibroblast skin	TruSeq Stranded Total RNA Library	Paired-end, 100 bp, 100 M	Progeria patient (HSS) (2 replicates)
			Paired-end, 75 bp, 100 M	Healthy donor, sex/age matched with progeria patient (2 replicates)
	Blood	TruSeq Stranded Total RNA Library	Paired-end, 75 bp, 70 M	Healthy donor, 6 females, 6 males, age range: 21-70

Supplementary Table 2.2: Description of the RNA-seq datasets used in the study.

² Not included in the analysis due to low coverage (below 1X genome coverage or 29 M sequenced reads for *R. norvegicus* ; genome coverage calculated using Lander-Waterman formula)

Bibliography

1. Kennedy, B. K. *et al.* Geroscience: Linking Aging to Chronic Disease. *Cell* **159**, 709–713 (2014).
2. Scott, A. J., Ellison, M. & Sinclair, D. A. The economic value of targeting aging. *Nat. Aging* **1**, 616–623 (2021).
3. Partridge, L., Deelen, J. & Slagboom, P. E. Facing up to the global challenges of ageing. *Nature* **561**, 45–56 (2018).
4. López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M. & Kroemer, G. The Hallmarks of Aging. *Cell* **153**, 1194–1217 (2013).
5. Vermulst, M. *et al.* Transcription errors induce proteotoxic stress and shorten cellular lifespan. *Nat. Commun.* **6**, 8065 (2015).
6. Martinez-Jimenez, C. P. *et al.* Aging increases cell-to-cell transcriptional variability upon immune stimulation. *Science* **355**, 1433–1436 (2017).
7. Saldi, T., Cortazar, M. A., Sheridan, R. M. & Bentley, D. L. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing. *J. Mol. Biol.* **428**, 2623–2635 (2016).
8. Bentley, D. L. Coupling mRNA processing with transcription in time and space. *Nat. Rev. Genet.* **15**, 163–175 (2014).
9. Aslanzadeh, V., Huang, Y., Sanguinetti, G. & Beggs, J. D. Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast. *Genome Res.* **28**, 203–213 (2018).
10. Ragan, C., Goodall, G. J., Shirokikh, N. E. & Preiss, T. Insights into the biogenesis and potential functions of exonic circular RNA. *Sci. Rep.* **9**, 2048 (2019).
11. van Leeuwen, F. W. *et al.* Frameshift Mutants of β Amyloid Precursor Protein and Ubiquitin-B in Alzheimer's and Down Patients. *Science* **279**, 242–247 (1998).
12. van Leeuwen, F. W., Hol, E. M. & Burbach, . Peter H. Mutations in RNA: a first example of molecular misreading in Alzheimer's disease. *Trends Neurosci.* **21**, 331–335 (1998).

13. Saxowsky, T. T., Meadows, K. L., Klungland, A. & Doetsch, P. W. 8-Oxoguanine-mediated transcriptional mutagenesis causes Ras activation in mammalian cells. *Proc. Natl. Acad. Sci.* **105**, 18877–18882 (2008).
14. Johnson, A., Alberts, B. & Lewis, J. *Molecular Biology of the Cell: A Problems Approach*. (Taylor & Francis, 2014).
15. Lee, T. I. & Young, R. A. Transcriptional Regulation and Its Misregulation in Disease. *Cell* **152**, 1237–1251 (2013).
16. Cramer, P. Eukaryotic Transcription Turns 50. *Cell* **179**, 808–812 (2019).
17. Hu, Z. *et al.* Nucleosome loss leads to global transcriptional up-regulation and genomic instability during yeast aging. *Genes Dev.* **28**, 396–408 (2014).
18. Stegeman, R. & Weake, V. M. Transcriptional Signatures of Aging. *J. Mol. Biol.* **429**, 2427–2437 (2017).
19. Kimmel, J. C. *et al.* Murine single-cell RNA-seq reveals cell-identity- and tissue-specific trajectories of aging. *Genome Res.* **29**, 2088–2103 (2019).
20. Benayoun, B. A. *et al.* Remodeling of epigenome and transcriptome landscapes with aging in mice reveals widespread induction of inflammatory responses. *Genome Res.* **29**, 697–709 (2019).
21. Sun, D. *et al.* Epigenomic profiling of young and aged HSCs reveals concerted changes during aging that reinforce self-renewal. *Cell Stem Cell* **14**, 673–688 (2014).
22. Schier, A. C. & Taatjes, D. J. Structure and mechanism of the RNA polymerase II transcription machinery. *Genes Dev.* **34**, 465–488 (2020).
23. Venkatesh, S. & Workman, J. L. Histone exchange, chromatin structure and the regulation of transcription. *Nat. Rev. Mol. Cell Biol.* **16**, 178–189 (2015).
24. Dollinger, R. & Gilmour, D. S. Regulation of Promoter Proximal Pausing of RNA Polymerase II in Metazoans. *J. Mol. Biol.* **433**, 166897 (2021).
25. Mayer, A., Landry, H. M. & Churchman, L. S. Pause & go: from the discovery of RNA

- polymerase pausing to its functional implications. *Curr. Opin. Cell Biol.* **46**, 72–80 (2017).
26. Wada, T. *et al.* DSIF, a novel transcription elongation factor that regulates RNA polymerase II processivity, is composed of human Spt4 and Spt5 homologs. *Genes Dev.* **12**, 343–356 (1998).
 27. Yamaguchi, Y. *et al.* NELF, a Multisubunit Complex Containing RD, Cooperates with DSIF to Repress RNA Polymerase II Elongation. *Cell* **97**, 41–51 (1999).
 28. Li, Y., Liu, M., Chen, L.-F. & Chen, R. P-TEFb: Finding its ways to release promoter-proximally paused RNA polymerase II. *Transcription* **9**, 88–94 (2018).
 29. Nudler, E. Transcription elongation: structural basis and mechanisms 11 Edited by M. Gottesman. *J. Mol. Biol.* **288**, 1–12 (1999).
 30. Brueckner, F. & Cramer, P. Structural basis of transcription inhibition by α -amanitin and implications for RNA polymerase II translocation. *Nat. Struct. Mol. Biol.* **15**, 811–818 (2008).
 31. Nudler, E. RNA Polymerase Backtracking in Gene Regulation and Genome Instability. *Cell* **149**, 1438–1445 (2012).
 32. Kamieniarz-Gdula, K. & Proudfoot, N. J. Transcriptional Control by Premature Termination: A Forgotten Mechanism. *Trends Genet.* **35**, 553–564 (2019).
 33. Lai, W. K. M. & Pugh, B. F. Understanding nucleosome dynamics and their links to gene expression and DNA replication. *Nat. Rev. Mol. Cell Biol.* **18**, 548–562 (2017).
 34. Liu, Y. *et al.* Phosphorylation of the Transcription Elongation Factor Spt5 by Yeast Bur1 Kinase Stimulates Recruitment of the PAF Complex. *Mol. Cell. Biol.* **29**, 4852–4863 (2009).
 35. Andrulis, E. D., Guzmán, E., Döring, P., Werner, J. & Lis, J. T. High-resolution localization of *Drosophila* Spt5 and Spt6 at heat shock genes in vivo: roles in promoter proximal pausing and transcription elongation. *Genes Dev.* **14**, 2635–2649 (2000).
 36. Vos, S. M., Farnung, L., Linden, A., Urlaub, H. & Cramer, P. Structure of complete Pol II–DSIF–PAF–SPT6 transcription complex reveals RTF1 allosteric activation. *Nat. Struct. Mol. Biol.* **27**, 668–677 (2020).
 37. Pruneski, J. A., Hainer, S. J., Petrov, K. O. & Martens, J. A. The Paf1 Complex Represses

SER3 Transcription in *Saccharomyces cerevisiae* by Facilitating Intergenic

Transcription-Dependent Nucleosome Occupancy of the SER3 Promoter. *Eukaryot. Cell* **10**, 1283–1294 (2011).

38. Wood, A., Schneider, J., Dover, J., Johnston, M. & Shilatifard, A. The Paf1 Complex Is Essential for Histone Monoubiquitination by the Rad6-Bre1 Complex, Which Signals for Histone Methylation by COMPASS and Dot1p*. *J. Biol. Chem.* **278**, 34739–34742 (2003).
39. Krogan, N. J. *et al.* The Paf1 Complex Is Required for Histone H3 Methylation by COMPASS and Dot1p: Linking Transcriptional Elongation to Histone Methylation. *Mol. Cell* **11**, 721–729 (2003).
40. Simic, R. *et al.* Chromatin remodeling protein Chd1 interacts with transcription elongation factors and localizes to transcribed genes. *EMBO J.* **22**, 1846–1856 (2003).
41. DeGennaro, C. M. *et al.* Spt6 Regulates Intragenic and Antisense Transcription, Nucleosome Positioning, and Histone Modifications Genome-Wide in Fission Yeast. *Mol. Cell. Biol.* **33**, 4779–4792 (2013).
42. Martin, B. J. E., Chruscicki, A. T. & Howe, L. J. Transcription Promotes the Interaction of the Facilitates Chromatin Transactions (FACT) Complex with Nucleosomes in *Saccharomyces cerevisiae*. *Genetics* **210**, 869–881 (2018).
43. Schwabish, M. A. & Struhl, K. Evidence for Eviction and Rapid Deposition of Histones upon Transcriptional Elongation by RNA Polymerase II. *Mol. Cell. Biol.* **24**, 10111–10117 (2004).
44. Proudfoot, N. J. Ending the message: poly(A) signals then and now. *Genes Dev.* **25**, 1770–1782 (2011).
45. Sun, Y. *et al.* Molecular basis for the recognition of the human AAUAAA polyadenylation signal. *Proc. Natl. Acad. Sci.* **115**, E1419–E1428 (2018).
46. Clerici, M., Faini, M., Aebersold, R. & Jinek, M. Structural insights into the assembly and polyA signal recognition mechanism of the human CPSF complex. *eLife* **6**, e33111 (2017).
47. Logan, J., Falck-Pedersen, E., Darnell, J. E. & Shenk, T. A poly(A) addition site and a

- downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. *Proc. Natl. Acad. Sci.* **84**, 8306–8310 (1987).
48. Connelly, S. & Manley, J. L. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev.* **2**, 440–452 (1988).
49. West, S., Gromak, N. & Proudfoot, N. J. Human 5' → 3' exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* **432**, 522–525 (2004).
50. Cortazar, M. A. *et al.* Control of RNA Pol II Speed by PNUTS-PP1 and Spt5 Dephosphorylation Facilitates Termination by a “Sitting Duck Torpedo” Mechanism. *Mol. Cell* **76**, 896-908.e4 (2019).
51. Sehgal, P. B., Derman, E., Molloy, G. R., Tamm, I. & Darnell, J. E. 5,6-Dichloro-1-β-D-Ribofuranosylbenzimidazole Inhibits Initiation of Nuclear Heterogeneous RNA Chains in HeLa Cells. *Science* **194**, 431–433 (1976).
52. Shermoen, A. W. & O'Farrell, P. H. Progression of the cell cycle through mitosis leads to abortion of nascent transcripts. *Cell* **67**, 303–310 (1991).
53. O'Brien, T. & Lis, J. T. Rapid changes in Drosophila transcription after an instantaneous heat shock. *Mol. Cell. Biol.* **13**, 3456–3463 (1993).
54. Tennyson, C. N., Klamut, H. J. & Worton, R. G. The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nat. Genet.* **9**, 184–190 (1995).
55. Mason, P. B. & Struhl, K. Distinction and Relationship between Elongation Rate and Processivity of RNA Polymerase II In Vivo. *Mol. Cell* **17**, 831–840 (2005).
56. Wada, Y. *et al.* A wave of nascent transcription on activated human genes. *Proc. Natl. Acad. Sci.* **106**, 18357–18361 (2009).
57. Singh, J. & Padgett, R. A. Rates of in situ transcription and splicing in large human genes. *Nat. Struct. Mol. Biol.* **16**, 1128–1133 (2009).
58. Darzacq, X. *et al.* In vivo dynamics of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.* **14**, 796–806 (2007).

59. Ben-Ari, Y. *et al.* The life of an mRNA in space and time. *J. Cell Sci.* **123**, 1761–1774 (2010).
60. Yunger, S., Rosenfeld, L., Garini, Y. & Shav-Tal, Y. Single-allele analysis of transcription kinetics in living mammalian cells. *Nat. Methods* **7**, 631–633 (2010).
61. Brody, Y. *et al.* The In Vivo Kinetics of RNA Polymerase II Elongation during Co-Transcriptional Splicing. *PLOS Biol.* **9**, e1000573 (2011).
62. Muramoto, T. *et al.* Live imaging of nascent RNA dynamics reveals distinct types of transcriptional pulse regulation. *Proc. Natl. Acad. Sci.* **109**, 7350–7355 (2012).
63. Maiuri, P. *et al.* Fast transcription rates of RNA polymerase II in human cells. *EMBO Rep.* **12**, 1280–1285 (2011).
64. Hocine, S., Raymond, P., Zenklusen, D., Chao, J. A. & Singer, R. H. Single-molecule analysis of gene expression using two-color RNA labeling in live yeast. *Nat. Methods* **10**, 119–121 (2013).
65. Fukaya, T., Lim, B. & Levine, M. Rapid Rates of Pol II Elongation in the Drosophila Embryo. *Curr. Biol.* **27**, 1387–1391 (2017).
66. Ardehali, M. B. *et al.* Spt6 enhances the elongation rate of RNA polymerase II in vivo. *EMBO J.* **28**, 1067–1077 (2009).
67. Yao, J., Ardehali, M. B., Fecko, C. J., Webb, W. W. & Lis, J. T. Intranuclear Distribution and Local Dynamics of RNA Polymerase II during Transcription Activation. *Mol. Cell* **28**, 978–990 (2007).
68. Collas, P. The Current State of Chromatin Immunoprecipitation. *Mol. Biotechnol.* **45**, 87–100 (2010).
69. Churchman, L. S. & Weissman, J. S. Nascent transcript sequencing visualizes transcription at nucleotide resolution. *Nature* **469**, 368–373 (2011).
70. Larson, M. H. *et al.* A pause sequence enriched at translation start sites drives transcription dynamics in vivo. *Science* **344**, 1042–1047 (2014).
71. Nojima, T. *et al.* Mammalian NET-Seq Reveals Genome-wide Nascent Transcription

- Coupled to RNA Processing. *Cell* **161**, 526–540 (2015).
72. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science* **322**, 1845–1848 (2008).
 73. Hah, N. *et al.* A Rapid, Extensive, and Transient Transcriptional Response to Estrogen Signaling in Breast Cancer Cells. *Cell* **145**, 622–634 (2011).
 74. Danko, C. G. *et al.* Signaling Pathways Differentially Affect RNA Polymerase II Initiation, Pausing, and Elongation Rate in Cells. *Mol. Cell* **50**, 212–222 (2013).
 75. Saponaro, M. *et al.* RECQL5 Controls Transcript Elongation and Suppresses Genome Instability Associated with Transcription Stress. *Cell* **157**, 1037–1049 (2014).
 76. Veloso, A. *et al.* Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* **24**, 896–905 (2014).
 77. Fuchs, G. *et al.* 4sUDRB-seq: measuring genomewide transcriptional elongation rates and initiation frequencies within cells. *Genome Biol.* **15**, R69 (2014).
 78. Fuchs, G. *et al.* Simultaneous measurement of genome-wide transcription elongation speeds and rates of RNA polymerase II transition into active elongation with 4sUDRB-seq. *Nat. Protoc.* **10**, 605–618 (2015).
 79. Baluapuri, A. *et al.* MYC Recruits SPT5 to RNA Polymerase II to Promote Processive Transcription Elongation. *Mol. Cell* **74**, 674-687.e11 (2019).
 80. Gregersen, L. H., Mitter, R. & Svejstrup, J. Q. Using TTchem-seq for profiling nascent transcription and measuring transcript elongation. *Nat. Protoc.* **15**, 604–627 (2020).
 81. Zhang, Y. *et al.* The Biogenesis of Nascent Circular RNAs. *Cell Rep.* **15**, 611–624 (2016).
 82. Ameur, A. *et al.* Total RNA sequencing reveals nascent transcription and widespread co-transcriptional splicing in the human brain. *Nat. Struct. Mol. Biol.* **18**, 1435–1440 (2011).
 83. Jonkers, I., Kwak, H. & Lis, J. T. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife* **3**, e02407 (2014).
 84. Caudron-Herger, M., Cook, P. R., Rippe, K. & Papantonis, A. Dissecting the nascent human

- transcriptome by analysing the RNA content of transcription factories. *Nucleic Acids Res.* **43**, e95 (2015).
85. Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* **389**, 251–260 (1997).
86. Szerlong, H. J. & Hansen, J. C. Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure. *Biochem. Cell Biol. Biochim. Biol. Cell.* **89**, 24–34 (2011).
87. Robinson, P. J. J., Fairall, L., Huynh, V. A. T. & Rhodes, D. EM measurements define the dimensions of the ‘30-nm’ chromatin fiber: evidence for a compact, interdigitated structure. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 6506–6511 (2006).
88. Bintu, L. *et al.* Nucleosomal Elements that Control the Topography of the Barrier to Transcription. *Cell* **151**, 738–749 (2012).
89. Weber, C. M., Ramachandran, S. & Henikoff, S. Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase. *Mol. Cell* **53**, 819–830 (2014).
90. Jimeno-González, S. *et al.* Defective histone supply causes changes in RNA polymerase II elongation rate and cotranscriptional pre-mRNA splicing. *Proc. Natl. Acad. Sci.* **112**, 14840–14845 (2015).
91. Sharma, A. *et al.* Calcium-mediated histone modifications regulate alternative splicing in cardiomyocytes. *Proc. Natl. Acad. Sci.* **111**, E4920–E4928 (2014).
92. Schor, I. E., Fiszbein, A., Petrillo, E. & Kornblihtt, A. R. Intragenic epigenetic changes modulate NCAM alternative splicing in neuronal differentiation. *EMBO J.* **32**, 2264–2274 (2013).
93. Fuchs, G., Hollander, D., Voichek, Y., Ast, G. & Oren, M. Cotranscriptional histone H2B monoubiquitylation is tightly coupled with RNA polymerase II elongation rate. *Genome Res.* **24**, 1572–1583 (2014).
94. Shukla, S. *et al.* CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature* **479**, 74–79 (2011).

95. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 167–177 (2015).
96. Cholewa-Waclaw, J. *et al.* Quantitative modelling predicts the impact of DNA methylation on RNA polymerase II traffic. *Proc. Natl. Acad. Sci.* **116**, 14995–15000 (2019).
97. Zamft, B., Bintu, L., Ishibashi, T. & Bustamante, C. Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 8948–8953 (2012).
98. Sigurdsson, S., Dirac-Svejstrup, A. B. & Svejstrup, J. Q. Evidence that Transcript Cleavage Is Essential for RNA Polymerase II Transcription and Cell Viability. *Mol. Cell* **38**, 202–210 (2010).
99. Cohen, E., Zafrir, Z. & Tuller, T. A code for transcription elongation speed. *RNA Biol.* **15**, 81–94 (2018).
100. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772–778 (2006).
101. Tilgner, H. *et al.* Nucleosome positioning as a determinant of exon recognition. *Nat. Struct. Mol. Biol.* **16**, 996–1001 (2009).
102. Spies, N., Nielsen, C. B., Padgett, R. A. & Burge, C. B. Biased Chromatin Signatures around Polyadenylation Sites and Exons. *Mol. Cell* **36**, 245–254 (2009).
103. Andersson, R., Enroth, S., Rada-Iglesias, A., Wadelius, C. & Komorowski, J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Res.* **19**, 1732–1741 (2009).
104. Harlen, K. M. *et al.* Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. *Cell Rep.* **15**, 2147–2158 (2016).
105. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing. *Science* **339**, 950–953 (2013).
106. Cheung, A. C. M. & Cramer, P. Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature* **471**, 249–253 (2011).
107. Kruk, J. A., Dutta, A., Fu, J., Gilmour, D. S. & Reese, J. C. The multifunctional Ccr4–Not

- complex directly promotes transcription elongation. *Genes Dev.* **25**, 581–593 (2011).
108. Sheridan, R. M., Fong, N., D'Alessandro, A. & Bentley, D. L. Widespread Backtracking by RNA Pol II Is a Major Effector of Gene Activation, 5' Pause Release, Termination, and Transcription Elongation Rate. *Mol. Cell* **73**, 107-118.e4 (2019).
109. Adkins, M. W. & Tyler, J. K. Transcriptional Activators Are Dispensable for Transcription in the Absence of Spt6-Mediated Chromatin Reassembly of Promoter Regions. *Mol. Cell* **21**, 405–416 (2006).
110. Quan, T. K. & Hartzog, G. A. Histone H3K4 and K36 Methylation, Chd1 and Rpd3S Oppose the Functions of *Saccharomyces cerevisiae* Spt4–Spt5 in Transcription. *Genetics* **184**, 321–334 (2010).
111. Fitz, J., Neumann, T. & Pavri, R. Regulation of RNA polymerase II processivity by Spt5 is restricted to a narrow window during elongation. *EMBO J.* **37**, e97965 (2018).
112. Liang, K. *et al.* Targeting Processive Transcription Elongation via SEC Disruption for MYC-Induced Cancer Therapy. *Cell* **175**, 766-779.e17 (2018).
113. García, A., Collin, A. & Calvo, O. Sub1 associates with Spt5 and influences RNA polymerase II transcription elongation rate. *Mol. Biol. Cell* **23**, 4297–4312 (2012).
114. Fan, Z. *et al.* CDK13 cooperates with CDK12 to control global RNA polymerase II processivity. *Sci. Adv.* **6**, eaaz5041 (2020).
115. Gregersen, L. H. *et al.* SCAF4 and SCAF8, mRNA Anti-Terminator Proteins. *Cell* **177**, 1797-1813.e18 (2019).
116. Hou, L. *et al.* Paf1C regulates RNA polymerase II progression by modulating elongation rate. *Proc. Natl. Acad. Sci.* **116**, 14583–14592 (2019).
117. Greenleaf, A. L., Borsett, L. M., Jiamachello, P. F. & Coulter, D. E. α -amanitin-resistant *D. melanogaster* with an altered RNA polymerase II. *Cell* **18**, 613–622 (1979).
118. de la Mata, M. *et al.* A Slow RNA Polymerase II Affects Alternative Splicing In Vivo. *Mol. Cell* **12**, 525–532 (2003).

119. Fong, N. *et al.* Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* **28**, 2663–2676 (2014).
120. Chow, L. T., Gelinas, R. E., Broker, T. R. & Roberts, R. J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **12**, 1–8 (1977).
121. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 3171–3175 (1977).
122. Dewey, C. N., Rogozin, I. B. & Koonin, E. V. Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns. *BMC Genomics* **7**, 311 (2006).
123. Will, C. L. & Lührmann, R. Spliceosome Structure and Function. *Cold Spring Harb. Perspect. Biol.* **3**, a003707 (2011).
124. Nieto Moreno, N., Giono, L. E., Cambindo Botto, A. E., Muñoz, M. J. & Kornblihtt, A. R. Chromatin, DNA structure and alternative splicing. *FEBS Lett.* **589**, 3370–3378 (2015).
125. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
126. Braberg, H. *et al.* From Structure to Systems: High-Resolution, Quantitative Genetic Analysis of RNA Polymerase II. *Cell* **154**, 775–788 (2013).
127. Khodor, Y. L. *et al.* Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in *Drosophila*. *Genes Dev.* **25**, 2502–2512 (2011).
128. Schor, I. E., Rascovan, N., Pelisch, F., Alló, M. & Kornblihtt, A. R. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. *Proc. Natl. Acad. Sci.* **106**, 4325–4330 (2009).
129. Ip, J. Y. *et al.* Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation. *Genome Res.* **21**, 390–401 (2011).
130. Dujardin, G. *et al.* How Slow RNA Polymerase II Elongation Favors Alternative Exon Skipping. *Mol. Cell* **54**, 683–690 (2014).

131. Maslon, M. M. *et al.* A slow transcription rate causes embryonic lethality and perturbs kinetic coupling of neuronal genes. *EMBO J.* **38**, e101244 (2019).
132. Leng, X. *et al.* Organismal benefits of transcription speed control at gene boundaries. *EMBO Rep.* **21**, e49315 (2020).
133. Donczew, R. & Hahn, S. Mechanistic Differences in Transcription Initiation at TATA-Less and TATA-Containing Promoters. *Mol. Cell. Biol.* **38**, e00448-17 (2017).
134. Moqtaderi, Z., Geisberg, J. V. & Struhl, K. Extensive Structural Differences of Closely Related 3' mRNA Isoforms: Links to Pab1 Binding and mRNA Stability. *Mol. Cell* **72**, 849-861.e6 (2018).
135. Chen, W. *et al.* Alternative Polyadenylation: Methods, Findings, and Impacts. *Genomics Proteomics Bioinformatics* **15**, 287–300 (2017).
136. Kaplan, C. D., Jin, H., Zhang, I. L. & Belyanin, A. Dissection of Pol II Trigger Loop Function and Pol II Activity–Dependent Control of Start Site Selection In Vivo. *PLOS Genet.* **8**, e1002627 (2012).
137. Rojas-Duran, M. F. & Gilbert, W. V. Alternative transcription start site selection leads to large differences in translation activity in yeast. *RNA* **18**, 2299–2305 (2012).
138. Li, J. *et al.* Kinetic Competition between Elongation Rate and Binding of NELF Controls Promoter-Proximal Pausing. *Mol. Cell* **50**, 711–722 (2013).
139. Fong, N. *et al.* Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition. *Mol. Cell* **60**, 256–267 (2015).
140. Soares, L. M. *et al.* Determinants of Histone H3K4 Methylation Patterns. *Mol. Cell* **68**, 773-785.e6 (2017).
141. Bar-Nahum, G. *et al.* A Ratchet Mechanism of Transcription Elongation and Its Control. *Cell* **120**, 183–193 (2005).
142. Kaplan, C. D. The architecture of RNA polymerase fidelity. *BMC Biol.* **8**, 85 (2010).

143. Loh, E., Salk, J. J. & Loeb, L. A. Optimization of DNA polymerase mutation rates during bacterial evolution. *Proc. Natl. Acad. Sci.* **107**, 1154–1159 (2010).
144. Saxowsky, T. T. & Doetsch, P. W. RNA Polymerase Encounters with DNA Damage: Transcription-Coupled Repair or Transcriptional Mutagenesis? *Chem. Rev.* **106**, 474–488 (2006).
145. Rosenberger, R. F. & Hilton, J. The frequency of transcriptional and translational errors at nonsense codons in the lacZ gene of Escherichia coli. *Mol. Gen. Genet. MGG* **191**, 207–212 (1983).
146. Blank, A., Gallant, J. A., Burgess, R. R. & Loeb, L. A. An RNA polymerase mutant with reduced accuracy of chain elongation. *Biochemistry* **25**, 5920–5928 (1986).
147. Vassylyev, D. G. *et al.* Structural basis for substrate loading in bacterial RNA polymerase. *Nature* **448**, 163–168 (2007).
148. Trigger loop of RNA polymerase is a positional, not acid–base, catalyst for both transcription and proofreading | PNAS. <https://www.pnas.org/doi/abs/10.1073/pnas.1702383114>.
149. Yuzenkova, Y. *et al.* Stepwise mechanism for transcription fidelity. *BMC Biol.* **8**, 54 (2010).
150. Sydow, J. F. & Cramer, P. RNA polymerase fidelity and transcriptional proofreading. *Curr. Opin. Struct. Biol.* **19**, 732–739 (2009).
151. Orlova, M., Newlands, J., Das, A., Goldfarb, A. & Borukhov, S. Intrinsic transcript cleavage activity of RNA polymerase. *Proc. Natl. Acad. Sci.* **92**, 4596–4600 (1995).
152. Zenkin, N., Yuzenkova, Y. & Severinov, K. Transcript-Assisted Transcriptional Proofreading. *Science* **313**, 518–520 (2006).
153. Mishanina, T. V., Palo, M. Z., Nayak, D., Mooney, R. A. & Landick, R. Trigger loop of RNA polymerase is a positional, not acid–base, catalyst for both transcription and proofreading. *Proc. Natl. Acad. Sci.* **114**, E5103–E5112 (2017).
154. Borukhov, S., Sagitov, V. & Goldfarb, A. Transcript cleavage factors from E. coli. *Cell* **72**, 459–466 (1993).
155. Sosunova, E. *et al.* Donation of catalytic residues to RNA polymerase active center by

- transcription factor Gre. *Proc. Natl. Acad. Sci.* **100**, 15469–15474 (2003).
156. Jeon, C. & Agarwal, K. Fidelity of RNA polymerase II transcription controlled by elongation factor TFIIS. *Proc. Natl. Acad. Sci.* **93**, 13677–13682 (1996).
157. Wons, E., Furmanek-Blaszczak, B. & Sektas, M. RNA editing by T7 RNA polymerase bypasses InDel mutations causing unexpected phenotypic changes. *Nucleic Acids Res.* **43**, 3950–3963 (2015).
158. Schwanhäusser, B. *et al.* Global quantification of mammalian gene expression control. *Nature* **473**, 337–342 (2011).
159. Ezerskyte, M. *et al.* O6-methylguanine-induced transcriptional mutagenesis reduces p53 tumor-suppressor function. *Proc. Natl. Acad. Sci.* **115**, 4731–4736 (2018).
160. Carey, L. B. RNA polymerase errors cause splicing defects and can be regulated by differential expression of RNA polymerase subunits. *eLife* **4**, e09945.
161. James, K., Gamba, P., Cockell, S. J. & Zenkin, N. Misincorporation by RNA polymerase is a major source of transcription pausing in vivo. *Nucleic Acids Res.* **45**, 1105–1113 (2017).
162. Gout, J.-F. *et al.* The landscape of transcription errors in eukaryotic cells. *Sci. Adv.* **3**, e1701484 (2017).
163. Springgate, C. F. & Loeb, L. A. On the fidelity of transcription by Escherichia coli ribonucleic acid polymerase. *J. Mol. Biol.* **97**, 577–591 (1975).
164. Rosenberger, R. F. & Foskett, G. An estimate of the frequency of in vivo transcriptional errors at a nonsense codon in Escherichia coli. *Mol. Gen. Genet. MGG* **183**, 561–563 (1981).
165. Roghanian, M., Zenkin, N. & Yuzenkova, Y. Bacterial global regulators DksA/ppGpp increase fidelity of transcription. *Nucleic Acids Res.* **43**, 1529–1536 (2015).
166. Satory, D. *et al.* DksA involvement in transcription fidelity buffers stochastic epigenetic change. *Nucleic Acids Res.* **43**, 10190–10199 (2015).
167. Shaw, R. J., Bonawitz, N. D. & Reines, D. Use of an in Vivo Reporter Assay to Test for Transcriptional and Translational Fidelity in Yeast. *J. Biol. Chem.* **277**, 24420–24426 (2002).

168. Nesser, N. K., Peterson, D. O. & Hawley, D. K. RNA polymerase II subunit Rpb9 is important for transcriptional fidelity in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 3268–3273 (2006).
169. Koyama, H., Ito, T., Nakanishi, T., Kawamura, N. & Sekimizu, K. Transcription elongation factor S-II maintains transcriptional fidelity and confers oxidative stress resistance. *Genes Cells Devoted Mol. Cell. Mech.* **8**, 779–788 (2003).
170. Bubunenko, M. G. *et al.* A Cre Transcription Fidelity Reporter Identifies GreA as a Major RNA Proofreading Factor in Escherichia coli. *Genetics* **206**, 179–187 (2017).
171. Irvin, J. D. *et al.* A Genetic Assay for Transcription Errors Reveals Multilayer Control of RNA Polymerase II Fidelity. *PLoS Genet.* **10**, e1004532 (2014).
172. Imashimizu, M., Oshima, T., Lubkowska, L. & Kashlev, M. Direct assessment of transcription fidelity by high-resolution RNA sequencing. *Nucleic Acids Res.* **41**, 9090–9104 (2013).
173. Loman, N. J. *et al.* Performance comparison of benchtop high-throughput sequencing platforms. *Nat. Biotechnol.* **30**, 434–439 (2012).
174. Meacham, F. *et al.* Identification and correction of systematic error in high-throughput sequence data. *BMC Bioinformatics* **12**, 451 (2011).
175. Gout, J.-F., Thomas, W. K., Smith, Z., Okamoto, K. & Lynch, M. Large-scale detection of in vivo transcription errors. *Proc. Natl. Acad. Sci.* **110**, 18584–18589 (2013).
176. Lou, D. I. *et al.* High-throughput DNA sequencing errors are reduced by orders of magnitude using circle sequencing. *Proc. Natl. Acad. Sci.* **110**, 19872–19877 (2013).
177. Acevedo, A., Brodsky, L. & Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505**, 686–690 (2014).
178. Traverse, C. C. & Ochman, H. Conserved rates and patterns of transcription errors across bacterial growth states and lifestyles. *Proc. Natl. Acad. Sci.* **113**, 3311–3316 (2016).
179. Reid-Bayliss, K. S. & Loeb, L. A. Accurate RNA consensus sequencing for high-fidelity detection of transcriptional mutagenesis-induced epimutations. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 9415–9420 (2017).

180. Imashimizu, M. *et al.* Visualizing translocation dynamics and nascent transcript errors in paused RNA polymerases in vivo. *Genome Biol.* **16**, 98 (2015).
181. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**, 377–382 (2009).
182. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
183. Picelli, S. *et al.* Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
184. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
185. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
186. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107.e17 (2018).
187. Hagemann-Jensen, M. *et al.* Single-cell RNA counting at allele and isoform resolution using Smart-seq3. *Nat. Biotechnol.* **38**, 708–714 (2020).
188. Hashimshony, T. *et al.* CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. *Genome Biol.* **17**, 77 (2016).
189. Kivioja, T. *et al.* Counting absolute numbers of molecules using unique molecular identifiers. *Nat. Methods* **9**, 72–74 (2012).
190. Hashimshony, T., Wagner, F., Sher, N. & Yanai, I. CEL-Seq: Single-Cell RNA-Seq by Multiplexed Linear Amplification. *Cell Rep.* **2**, 666–673 (2012).
191. Sasagawa, Y. *et al.* Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biol.* **19**, 29 (2018).
192. Fu, G. K., Hu, J., Wang, P.-H. & Fodor, S. P. A. Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proc. Natl. Acad. Sci.* **108**, 9026–9031 (2011).

193. Casbon, J. A., Osborne, R. J., Brenner, S. & Lichtenstein, C. P. A method for counting PCR template molecules with application to next-generation sequencing. *Nucleic Acids Res.* **39**, e81 (2011).
194. Kukita, Y. *et al.* High-fidelity target sequencing of individual molecules identified using barcode sequences: de novo detection and absolute quantitation of mutations in plasma cell-free DNA from cancer patients. *DNA Res. Int. J. Rapid Publ. Rep. Genes Genomes* **22**, 269–277 (2015).
195. Schmitt, M. W. *et al.* Detection of ultra-rare mutations by next-generation sequencing. *Proc. Natl. Acad. Sci.* **109**, 14508–14513 (2012).
196. Orton, R. J. *et al.* Distinguishing low frequency mutations from RT-PCR and sequence errors in viral deep sequencing data. *BMC Genomics* **16**, 229 (2015).
197. Fu, Y., Wu, P.-H., Beane, T., Zamore, P. D. & Weng, Z. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BMC Genomics* **19**, 531 (2018).
198. Sena, J. A. *et al.* Unique Molecular Identifiers reveal a novel sequencing artefact with implications for RNA-Seq based gene expression analysis. *Sci. Rep.* **8**, 13121 (2018).
199. Darnell, J. E. Variety in the level of gene control in eukaryotic cells. *Nature* **297**, 365–371 (1982).
200. Vogel, C. & Marcotte, E. M. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nat. Rev. Genet.* **13**, 227–232 (2012).
201. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535–550 (2016).
202. Conaway, J. W. & Conaway, R. C. Transcription Elongation and Human Disease. *Annu. Rev. Biochem.* **68**, 301–319 (1999).
203. Fritsch, C. *et al.* Genome-wide surveillance of transcription errors in response to genotoxic stress. *Proc. Natl. Acad. Sci. U. S. A.* **118**, e2004077118 (2021).
204. Rangaraju, S. *et al.* Suppression of transcriptional drift extends *C. elegans* lifespan by

- postponing the onset of mortality. *eLife* **4**, e08833 (2015).
205. Rogalski, T. M., Bullerjahn, A. M. & Riddle, D. L. Lethal and amanitin-resistance mutations in the *Caenorhabditis elegans* *ama-1* and *ama-2* genes. *Genetics* **120**, 409–422 (1988).
206. Chen, Y., Chafin, D., Price, D. H. & Greenleaf, A. L. *Drosophila* RNA Polymerase II Mutants That Affect Transcription Elongation (*). *J. Biol. Chem.* **271**, 5993–5999 (1996).
207. Oesterreich, F. C. *et al.* Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell* **165**, 372–381 (2016).
208. Braunschweig, U. *et al.* Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Res.* **24**, 1774–1786 (2014).
209. Jung, H. *et al.* Intron retention is a widespread mechanism of tumor-suppressor inactivation. *Nat. Genet.* **47**, 1242–1248 (2015).
210. Drexler, H. L., Choquet, K. & Churchman, L. S. Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Mol. Cell* **77**, 985-998.e8 (2020).
211. Mazin, P. *et al.* Widespread splicing changes in human brain development and aging. *Mol. Syst. Biol.* **9**, 633 (2013).
212. Tollervey, J. R. *et al.* Analysis of alternative splicing associated with aging and neurodegeneration in the human brain. *Genome Res.* **21**, 1572–1582 (2011).
213. Lee, B. P. *et al.* Changes in the expression of splicing factor transcripts and variations in alternative splicing are associated with lifespan in mice and humans. *Aging Cell* **15**, 903–913 (2016).
214. Heintz, C. *et al.* Splicing factor 1 modulates dietary restriction and TORC1 pathway longevity in *C. elegans*. *Nature* **541**, 102–106 (2017).
215. Pickrell, J. K., Pai, A. A., Gilad, Y. & Pritchard, J. K. Noisy Splicing Drives mRNA Isoform Diversity in Human Cells. *PLoS Genet.* **6**, e1001236 (2010).
216. Stepankiw, N., Raghavan, M., Fogarty, E. A., Grimson, A. & Pleiss, J. A. Widespread alternative and aberrant splicing revealed by lariat sequencing. *Nucleic Acids Res.* **43**,

- 8488–8501 (2015).
217. Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nat. Genet.* **50**, 151–158 (2018).
218. Deterioration of the human transcriptome with age due to increasing intron retention and spurious splicing | bioRxiv. <https://www.biorxiv.org/content/10.1101/2022.03.14.484341v1>.
219. Cocquerelle, C., Mascrez, B., Héтуin, D. & Bailleul, B. Mis-splicing yields circular RNA molecules. *FASEB J.* **7**, 155–160 (1993).
220. Nigro, J. M. *et al.* Scrambled exons. *Cell* **64**, 607–613 (1991).
221. Zhang, X.-O. *et al.* Complementary Sequence-Mediated Exon Circularization. *Cell* **159**, 134–147 (2014).
222. Feser, J. *et al.* Elevated Histone Expression Promotes Life Span Extension. *Mol. Cell* **39**, 724–735 (2010).
223. Hughes, A. L. & Rando, O. J. Mechanisms Underlying Nucleosome Positioning In Vivo. *Annu. Rev. Biophys.* **43**, 41–63 (2014).
224. Struhl, K. & Segal, E. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.* **20**, 267–273 (2013).
225. Fitz, V. *et al.* Nucleosomal arrangement affects single-molecule transcription dynamics. *Proc. Natl. Acad. Sci.* **113**, 12733–12738 (2016).
226. Gossett, A. J. & Lieb, J. D. In Vivo Effects of Histone H3 Depletion on Nucleosome Occupancy and Position in *Saccharomyces cerevisiae*. *PLOS Genet.* **8**, e1002771 (2012).
227. Oberdoerffer, P. An age of fewer histones. *Nat. Cell Biol.* **12**, 1029–1031 (2010).
228. Sural, S., Liang, C.-Y., Wang, F.-Y., Ching, T.-T. & Hsu, A.-L. HSB-1/HSF-1 pathway modulates histone H4 in mitochondria to control mtDNA transcription and longevity. *Sci. Adv.* **6**, eaaz4452 (2020).
229. Lu, Y.-X. *et al.* A TORC1-histone axis regulates chromatin organisation and non-canonical induction of autophagy to ameliorate ageing. *eLife* **10**, e62233 (2021).

230. Bushnell, D. A., Cramer, P. & Kornberg, R. D. Structural basis of transcription: α -Amanitin–RNA polymerase II cocystal at 2.8 Å resolution. *Proc. Natl. Acad. Sci.* **99**, 1218–1222 (2002).
231. Bowman, E. A., Riddle, D. L. & Kelly, W. Amino Acid Substitutions in the *Caenorhabditis elegans* RNA Polymerase II Large Subunit AMA-1/RPB-1 that Result in α -Amanitin Resistance and/or Reduced Function. *G3 GenesGenomesGenetics* **1**, 411–416 (2011).
232. Stroustrup, N. *et al.* The *Caenorhabditis elegans* Lifespan Machine. *Nat. Methods* **10**, 665–670 (2013).
233. Greenleaf, A. L., Borsett, L. M., Jiamachello, P. F. & Coulter, D. E. Alpha-amanitin-resistant *D. melanogaster* with an altered RNA polymerase II. *Cell* **18**, 613–622 (1979).
234. Grönke, S., Clarke, D.-F., Broughton, S., Andrews, T. D. & Partridge, L. Molecular evolution and functional characterization of *Drosophila* insulin-like peptides. *PLoS Genet.* **6**, e1000857 (2010).
235. Weigelt, C. M. *et al.* An Insulin-Sensitive Circular RNA that Regulates Lifespan in *Drosophila*. *Mol. Cell* **79**, 268-279.e5 (2020).
236. Hahn, O. *et al.* Dietary restriction protects from age-associated DNA methylation and induces epigenetic reprogramming of lipid metabolism. *Genome Biol.* **18**, 56 (2017).
237. Selman, C. *et al.* Evidence for lifespan extension and delayed age–related biomarkers in insulin receptor substrate 1 null mice. *FASEB J.* **22**, 807–818 (2008).
238. Melnik, S. *et al.* Isolation of the protein and RNA content of active sites of transcription from mammalian cells. *Nat. Protoc.* **11**, 553–565 (2016).
239. Lusser, A. *et al.* Thiouridine-to-Cytidine Conversion Sequencing (TUC-Seq) to Measure mRNA Transcription and Degradation Rates. in *The Eukaryotic RNA Exosome: Methods and Protocols* (eds. LaCava, J. & Vaňáčová, Š.) 191–211 (Springer, 2020).
doi:10.1007/978-1-4939-9822-7_10.
240. Essers, P. *et al.* Reduced insulin/insulin-like growth factor signaling decreases translation in

- Drosophila and mice. *Sci. Rep.* **6**, 30290 (2016).
241. Diermeier, S. *et al.* TNF α signalling primes chromatin for NF- κ B binding and induces rapid and widespread nucleosome repositioning. *Genome Biol.* **15**, 536 (2014).
242. Zhao, S. *et al.* PiggyBac transposon vectors: the tools of the human gene encoding. *Transl. Lung Cancer Res.* **5**, 120–125 (2016).
243. Adachi, K. *et al.* Esrrb Unlocks Silenced Enhancers for Reprogramming to Naive Pluripotency. *Cell Stem Cell* **23**, 266-275.e6 (2018).
244. Dimri, G. P. *et al.* A biomarker that identifies senescent human cells in culture and in aging skin in vivo. *Proc. Natl. Acad. Sci. U. S. A.* **92**, 9363–9367 (1995).
245. Zirkel, A. *et al.* HMGB2 Loss upon Senescence Entry Disrupts Genomic Organization and Induces CTCF Clustering across Cell Types. *Mol. Cell* **70**, 730-744.e6 (2018).
246. Berridge, M. V. & Tan, A. S. Characterization of the cellular reduction of 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT): subcellular localization, substrate dependence, and involvement of mitochondrial electron transport in MTT reduction. *Arch. Biochem. Biophys.* **303**, 474–482 (1993).
247. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
248. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).
249. Karolchik, D. *et al.* The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.* **32**, D493-496 (2004).
250. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
251. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
252. Gerstner, J. R. *et al.* Removal of unwanted variation reveals novel patterns of gene

- expression linked to sleep homeostasis in murine cortex. *BMC Genomics* **17**, 727 (2016).
253. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).
254. Durinck, S. *et al.* BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**, 3439–3440 (2005).
255. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
256. Hwang, T. *et al.* Dynamic regulation of RNA editing in human brain development and disease. *Nat. Neurosci.* **19**, 1093–1099 (2016).
257. Flores, O. & Orozco, M. nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics* **27**, 2149–2150 (2011).
258. Saldi, T., Riemondy, K., Erickson, B. & Bentley, D. L. Alternative RNA structures formed during transcription depend on elongation rate and modify RNA processing. *Mol. Cell* **81**, 1789-1801.e5 (2021).
259. Miller, T. E. *et al.* Transcription elongation factors represent in vivo cancer dependencies in glioblastoma. *Nature* **547**, 355–359 (2017).
260. Ogle, J. M. & Ramakrishnan, V. Structural Insights into Translational Fidelity. *Annu. Rev. Biochem.* **74**, 129–177 (2005).
261. Bacher, J. M., de Crécy-Lagard, V. & Schimmel, P. R. Inhibited cell growth and protein functional changes from an editing-defective tRNA synthetase. *Proc. Natl. Acad. Sci.* **102**, 1697–1701 (2005).
262. Drummond, D. A. & Wilke, C. O. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell* **134**, 341–352 (2008).
263. Lee, J. W. *et al.* Editing-defective tRNA synthetase causes protein misfolding and neurodegeneration. *Nature* **443**, 50–55 (2006).

264. Loeb, L. A. & Monnat, R. J. DNA polymerases and human disease. *Nat. Rev. Genet.* **9**, 594–604 (2008).
265. Fredriksson, Å. *et al.* Decline in ribosomal fidelity contributes to the accumulation and stabilization of the master stress response regulator σ S upon carbon starvation. *Genes Dev.* **21**, 862–874 (2007).
266. Brégeon, D. & Doetsch, P. W. Transcriptional mutagenesis: causes and involvement in tumour development. *Nat. Rev. Cancer* **11**, 218–227 (2011).
267. Gordon, A. J. E., Satory, D., Halliday, J. A. & Herman, C. Heritable Change Caused by Transient Transcription Errors. *PLOS Genet.* **9**, e1003595 (2013).
268. Paoloni-Giacobino, A., Rossier, C., Papasavvas, M. & Antonarakis, S. Frequency of replication/transcription errors in (A)/(T) runs of human genes. *Hum. Genet.* **109**, 40–47 (2001).
269. Strathern, J. N., Jin, D. J., Court, D. L. & Kashlev, M. Isolation and characterization of transcription fidelity mutants. *Biochim. Biophys. Acta BBA - Gene Regul. Mech.* **1819**, 694–699 (2012).
270. Ji, J. P. & Loeb, L. A. Fidelity of HIV-1 reverse transcriptase copying RNA in vitro. *Biochemistry* **31**, 954–958 (1992).
271. Dal Molin, A. & Di Camillo, B. How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives. *Brief. Bioinform.* **20**, 1384–1394 (2019).
272. Cao, Y. *et al.* Integrated analysis of multimodal single-cell data with structural similarity. *Nucleic Acids Res.* gkac781 (2022) doi:10.1093/nar/gkac781.
273. Almanzar, N. *et al.* A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590–595 (2020).
274. Savva, Y. A., Rieder, L. E. & Reenan, R. A. The ADAR protein family. *Genome Biol.* **13**, 252 (2012).
275. Nishikura, K. Functions and Regulation of RNA Editing by ADAR Deaminases. *Annu. Rev. Biochem.* **79**, 321–349 (2010).

276. Liu, H. *et al.* Functional Impact of RNA editing and ADARs on regulation of gene expression: perspectives from deep sequencing studies. *Cell Biosci.* **4**, 44 (2014).
277. Alseth, I., Dalhus, B. & Bjørås, M. Inosine in DNA and RNA. *Curr. Opin. Genet. Dev.* **26**, 116–123 (2014).
278. See, P., Lum, J., Chen, J. & Ginhoux, F. A Single-Cell Sequencing Guide for Immunologists. *Front. Immunol.* **9**, (2018).
279. Chung, C. *et al.* The fidelity of transcription in human cells. 2022.05.10.491385 Preprint at <https://doi.org/10.1101/2022.05.10.491385> (2022).
280. Kargapolova, Y., Levin, M., Lackner, K. & Danckwardt, S. sCLIP—an integrated platform to study RNA–protein interactomes in biomedical research: identification of CSTF2tau in alternative processing of small nuclear RNAs. *Nucleic Acids Res.* **45**, 6074–6086 (2017).
281. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. 2021.05.05.442755 Preprint at <https://doi.org/10.1101/2021.05.05.442755> (2021).
282. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
283. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
284. Rooney, J. *et al.* PCR Based Determination of Mitochondrial DNA Copy Number in Multiple Species. *Methods Mol. Biol. Clifton NJ* **1241**, 23–38 (2015).
285. Picardi, E., D’Erchia, A. M., Lo Giudice, C. & Pesole, G. REDportal: a comprehensive database of A-to-I RNA editing events in humans. *Nucleic Acids Res.* **45**, D750–D757 (2017).
286. Lawrence, M., Gentleman, R. & Carey, V. rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* **25**, 1841–1842 (2009).
287. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLOS Comput. Biol.* **9**, e1003118 (2013).

288. Menéndez-Arias, L. HIV Reverse Transcriptase Fidelity, Clade Diversity, and Acquisition of Drug Resistance. in *Human Immunodeficiency Virus Reverse Transcriptase* (eds. LeGrice, S. & Gotte, M.) 225–252 (Springer New York, 2013). doi:10.1007/978-1-4614-7291-9_11.
289. Yao, G., Lee, T. J., Mori, S., Nevins, J. R. & You, L. A bistable Rb–E2F switch underlies the restriction point. *Nat. Cell Biol.* **10**, 476–482 (2008).
290. Kotsantis, P. *et al.* Increased global transcription activity as a mechanism of replication stress in cancer. *Nat. Commun.* **7**, 13087 (2016).
291. Villeponteau, B. The heterochromatin loss model of aging. *Exp. Gerontol.* **32**, 383–394 (1997).
292. Tsurumi, A. & Li, W. Global heterochromatin loss. *Epigenetics* **7**, 680–688 (2012).
293. Ni, Z., Ebata, A., Alipanahramandi, E. & Lee, S. S. Two SET domain containing genes link epigenetic changes and aging in *Caenorhabditis elegans*. *Aging Cell* **11**, 315–325 (2012).
294. Ivanov, A. *et al.* Lysosome-mediated processing of chromatin in senescence. *J. Cell Biol.* **202**, 129–143 (2013).
295. O’Sullivan, R. J., Kubicek, S., Schreiber, S. L. & Karlseder, J. Reduced histone biosynthesis and chromatin changes arising from a damage signal at telomeres. *Nat. Struct. Mol. Biol.* **17**, 1218–1225 (2010).
296. McColl, G. *et al.* Pharmacogenetic Analysis of Lithium-induced Delayed Aging in *Caenorhabditis elegans**. *J. Biol. Chem.* **283**, 350–357 (2008).
297. Bochkis, I. M., Przybylski, D., Chen, J. & Regev, A. Changes in nucleosome occupancy associated with metabolic alterations in aged mammalian liver. *Cell Rep.* **9**, 996–1006 (2014).
298. Chen, Y., Bravo, J. I., Son, J. M., Lee, C. & Benayoun, B. A. Remodeling of the H3 nucleosomal landscape during mouse aging. *Transl. Med. Aging* **4**, 22–31 (2020).
299. Hofmann, J. W. *et al.* Reduced Expression of MYC Increases Longevity and Enhances Healthspan. *Cell* **160**, 477–488 (2015).
300. Lozano-Vidal, N. *et al.* The PNUTS-PP1 axis regulates endothelial aging and barrier function

via SEMA3B suppression. 2020.08.10.243170 Preprint at

<https://doi.org/10.1101/2020.08.10.243170> (2020).

301. Bozukova, M. *et al.* Aging is associated with increased chromatin accessibility and reduced polymerase pausing in liver. *Mol. Syst. Biol.* **18**, e11002 (2022).
302. Moore, J. E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
303. Zhao, C., Liu, F. & Pyle, A. M. An ultraprocessive, accurate reverse transcriptase encoded by a metazoan group II intron. *RNA N. Y. N* **24**, 183–195 (2018).
304. Zhou, S., Jones, C., Mieczkowski, P. & Swanstrom, R. Primer ID Validates Template Sampling Depth and Greatly Reduces the Error Rate of Next-Generation Sequencing of HIV-1 Genomic RNA Populations. *J. Virol.* **89**, 8540–8555 (2015).
305. Yasukawa, K. *et al.* Next-generation sequencing-based analysis of reverse transcriptase fidelity. *Biochem. Biophys. Res. Commun.* **492**, 147–153 (2017).
306. Álvarez, M. & Menéndez-Arias, L. Temperature effects on the fidelity of a thermostable HIV-1 reverse transcriptase. *FEBS J.* **281**, 342–351 (2014).
307. Rodriguez, J., Menet, J. S. & Rosbash, M. Nascent-seq indicates widespread cotranscriptional RNA editing in *Drosophila*. *Mol. Cell* **47**, 27–37 (2012).
308. Penn, A. C., Balik, A. & Greger, I. H. Steric antisense inhibition of AMPA receptor Q/R editing reveals tight coupling to intronic editing sites and splicing. *Nucleic Acids Res.* **41**, 1113–1123 (2013).
309. Rieder, L. E. & Reenan, R. A. The intricate relationship between RNA structure, editing, and splicing. *Semin. Cell Dev. Biol.* **23**, 281–288 (2012).
310. Lu, T. *et al.* Gene regulation and DNA damage in the ageing human brain. *Nature* **429**, 883–891 (2004).
311. Sedelnikova, O. A. *et al.* Role of oxidatively induced DNA lesions in human pathogenesis. *Mutat. Res. Mutat. Res.* **704**, 152–159 (2010).

312. Floyd, R. A. & Hensley, K. Oxidative stress in brain aging: Implications for therapeutics of neurodegenerative diseases. *Neurobiol. Aging* **23**, 795–807 (2002).
313. Hahm, J. Y., Park, J., Jang, E.-S. & Chi, S. W. 8-Oxoguanine: from oxidative damage to epigenetic and epitranscriptional modification. *Exp. Mol. Med.* **54**, 1626–1642 (2022).
314. Brégeon, D., Peignon, P.-A. & Sarasin, A. Transcriptional Mutagenesis Induced by 8-Oxoguanine in Mammalian Cells. *PLOS Genet.* **5**, e1000577 (2009).
315. Paredes, J. A., Ezerskyte, M., Bottai, M. & Dreij, K. Transcriptional mutagenesis reduces splicing fidelity in mammalian cells. *Nucleic Acids Res.* **45**, 6520–6529 (2017).
316. Basu, S., Je, G. & Kim, Y.-S. Transcriptional mutagenesis by 8-oxodG in α -synuclein aggregation and the pathogenesis of Parkinson's disease. *Exp. Mol. Med.* **47**, e179–e179 (2015).
317. Lareau, C. A. *et al.* Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).

Curriculum vitae

Antonios Papadakis

Personal Details

Date of Birth: August 14, 1991

Nationality: Greek

Address: Furvelser Str. 15,
51069, Cologne, Germany

E-mail: apapada1@uni-koeln.de

ORCID iD: 0000-0002-2416-6772

Research Experience

*Cologne Excellence Cluster on Cellular Stress Responses in Aging
Associated Diseases*

(CECAD), Cologne, Germany

PhD candidate, Cellular Networks and Systems Biology Group (March
2017 - present)

“Aging-associated changes of transcriptional elongation speed and
transcriptional error rate”

Biology Department, University of Crete, Heraklion, Greece

Master’s student, Computational Genomics Group (January 2016-October
2016)

“Computational Prioritization of Differential Expression Gene Lists through
Analysis of Regulatory and Functional Networks.”

Education

**Faculty of Mathematics and Natural Science, University of Cologne,
Germany**

PhD student(2018-present)

Biology Department, University of Crete, Heraklion, Greece

Master of Science in Molecular Biology and Biomedicine (2013-2016)

Grade (9.19/10)

Bachelor's in Biology (2009-2013)

Grade (8.4/10)

Selected scientific publications and conferences

Debès, C., **Papadakis, A.**, Grönke, S., Karalay, Ö., Tain, L., Nakamura, S., Hahn, O., ... & Beyer, A. (2022). Aging-associated changes in transcriptional elongation influence metazoan longevity. *bioRxiv*, 719864.

Hagmann, H., Khayyat, N. H., Oezel, C., **Papadakis, A.**, Kuczkowski, A., Benzing, T., ... & Brinkkoetter, P. T. (2022). Paraoxonase 2 (PON2) Deficiency Reproduces Lipid Alterations of Diabetic and Inflammatory Glomerular Disease and Affects TRPC6 Signaling. *Cells*, 11(22), 3625.

CMMC Annual Retreat 2021, Cologne, Germany (Feb 2021) – Poster presentation

ISMB/ECCB 2019, Basel, Switzerland (Jul 2019) – Poster presentation

Arbeitsgemeinschaft für Gen-Diagnostik annual meeting, Potsdam, Germany (Oct 2018)

Annual Meeting of the German Foundation for Aging Research (DGfA), Cologne, Germany, (Dec 2017)

Training & Awards

Good Research Practice workshop, University of Zurich, Zurich, Switzerland (Nov 2021)

1st poster prize winner, CMMC Annual Retreat 2021, Cologne, Germany (Feb 2021)

de.NBI Summer School: Computational Genomics and RNA Biology, Berlin, Germany(Sep. 2017)

Fellowship from the Greek Fellowship Foundation(2009-2011)

Skills

Programming Languages: R - daily usage | Python, Perl and Javascript–basic programming skills | Unix shell

Data Analysis: Single-cell RNA-seq/ATAC-seq | Bulk RNA-seq | MNase-seq | NET-seq| Network Topology Analysis

Development: Git | GitHub

Languages: Greek (Native) | English (C2) | French (C2) | German (Basic)