Mathematical model of transcription loss due to accumulated DNA damage

Marko Raseta*[1], Shannon Dealy[1], Jacinta van de Grint[1], Jiang Chang[2], Jan Hoeijmakers[1], and Joris Pothof[1]

[1]Department of Molecular Genetics, Erasmus MC, Netherlands
[2]University of Oxford, UK

Abstract

We offer a simple mathematical model of gene transcription loss due to accumulated DNA damage in time based on widely agreed biological axioms. Closed form formulae characterizing the distribution of the underlying stochastic processes representing the transcription loss upon specified number of DNA damages are obtained. Moreover, the asymptotic behavior of the stochastic process was analyzed. Finally, the distribution of the first hitting time of transcription loss to specified biologically relevant levels was studied both analytically and computationally on mice data.

Keywords: DNA damage, Transcription loss, Moment formulae, Limit behavior, Computational inference

Mathematics Subject Classification 92-10

1 Introduction

Transcription is the process of creating an RNA copy of a gene's DNA sequence called messenger RNA (mRNA) which represents a carrier of the gene's protein information encoded in DNA. In humans and other complex biological organisms, mRNA moves from the cell nucleus to the cell cytoplasm, where it is subsequently used in the process of synthesis the encoded protein. DNA damage represents any alteration of the chemical structure of the DNA molecule and can occur both naturally and due to presence of the exogenous factors. Moreover, DNA damages are changes in the

*m.raseta@erasmusmc.nl

structure of the genetic material and can prevent the transcription machinery from functioning and performing properly, causing transcription stress ([24]). Ageing is a naturally occurring biological process associated with a gradual decline in biological function. There is a growing body of scientific evidence that aging is a direct consequence of the accumulation of unrepaired DNA damage. This idea was first suggested in ([1]) with the ever increasing experimental proof over the past decades ([2, 3, 4, 5, 6, 7]). The accumulation of DNA damage is particularly visible in cells that are either non-replicating or slowly replicating, because DNA repair capacity is lower in these cells ([25]). This includes but is not restricted to cells in the brain ([8]), muscle ([9, 10, 11]), liver ([9, 12]) and kidney ([9, 13]). The corresponding reduction in gene expression is observed both on mRNA and protein levels. Further support of the DNA damage theory of aging comes from the observed accelerated aging in humans with inherited defects in DNA repair mechanisms, such as Werner syndrome ([14]), Huchinsosn-Gilford progeria ([15]) and Cockayne syndrome ([16]) with corresponding mean life expectancy of 47, 13 and 13 years, respectively. Moreover, it was recently demonstrated that the age-related transcriptional stress is evolutionary conserved from nematodes to humans. Thus, accumulation of stochastic endogenous DNA damage during aging deteriorates basal transcription, which establishes the age-related transcriptome and causes dysfunction of key aging hallmark pathways, disclosing how DNA damage functionally underlies major aspects of normal aging ([23]).

The purpose of this study is to quantify the connection between the accumulation of unrepaired DNA damage and the associated loss in transcription in a mathematically rigorous fashion and to study the stochastic properties of the associated processes. To be more specific, we quantify the distribution of the transcription loss as a function of the number of accumulated DNA damages in closed form. Furthermore, although the closed form formula for the distribution of the first hitting time of biologically relevant levels of blocked transcription is available, it is impractical due to computational complexity and hence simulation is used to draw inferences on these important distributions. Moreover, we also quantify the distribution of the number of damages needed to switch off both copies of the gene in the genome thereby terminating its biological function.

This manuscript is organized as follows. In Section 2 we list and corroborate the key biological assumptions of the model that are in turn translated into mathematical foundations. Section 3 presents the findings including closed form formulae for the distribution of losses as a function of number of DNA damages accumulated. Section 4 presents findings of a simulation study for the distribution of the first hitting time of the level of accumulated DNA damage to a range of biologically relevant levels. Finally, Section 5 provides some concluding remarks and elaborates on future research directions.

2 The set-up

We introduce a mathematical model of transcription loss due to accumulation of DNA damage in biological organisms. DNA damage accumulates at a constant discrete rate ([17, 18]). We shall assume that once certain gene exhibits a damage this damage is indeed permanent and the amount of transcription associated with the gene drops to 0 from that moment onward. There exist two DNA strands, one from each parent while the transcription is only relevant in one direction ([19]). We assume that damage is equally likely to occur on any one of these and hence, at any one time, with probability $\frac{1}{2}$, no transcription is lost. Each of the genes will have two copies and the transcription will be equally split between these. The probability that gene i will exhibit DNA damage is only assumed to be proportional to its length. Equivalently, we assume DNA is uniformly distributed over the entire length of the genome ([20, 21]).

Consider a biological organism with $2N$ genes in total, thereby taking into account presence of a copy of each gene. For all $i \in \{1, \dots, N\}$ let l_i and α_i stand for the length and weight of gene i, respectively, with restriction:

$$\sum_{i=1}^{2N} \alpha_i = 1 \tag{2.1}$$

where, for convenience, we rearranged the genome to have $\alpha_i = \alpha_{N+i}$ for all $i \in \{1, \dots, N\}$. In line with the assumptions above we will assume that, at any instance of time and independently of both past and future, the probability that gene i will exhibit DNA damage equals:

$$q_i = \frac{l_i}{4L}, L = \sum_{i=1}^{N} l_i \tag{2.2}$$

where again $p_i = p_{N+i}$ for all $i \in \{1, \dots, N\}$.

Let us introduce a partition of the interval $[0, 1]$ which we will use throughout the manuscript:

Definition 2.1. *For $i \in \{1, \dots, 2N+1\}$ define the following sequence of intervals*
$I_1 = [0, \frac{l_1}{4L})$
$I_2 = [\frac{l_1}{4L}, \frac{l_1+l_2}{4L})$
.
.

.

$$I_N = \left[\frac{l_1+l_2+\cdots+l_{N-1}}{4L}, \frac{1}{4}\right)$$

$$I_{N+1} = \left[\frac{1}{4}, \frac{1}{4} + \frac{l_1}{4L}\right)$$

$$I_{N+2} = \left[\frac{1}{4} + \frac{l_1}{4L}, \frac{1}{4} + \frac{l_1+l_2}{4L}\right)$$

.

.

.

$$I_{2N} = \left[\frac{1}{4} + \frac{l_1+l_2+\cdots+l_{N-1}}{4L}, \frac{1}{2}\right)$$

$$I_{2N+1} = [\frac{1}{2}, 1]$$

3 Results

We begin with a trivial yet important observation.

Proposition 1. *Let U be a uniformly distributed random variable on the interval $[0,1]$. Then:*
(i) $\mathbb{P}$ (gene i exhibits damage at any time) $= \mathbb{P}\ (U \in I_i)$
(ii) $\mathbb{P}$ (either one of the two inactive strands is hit at any one time)$= \mathbb{P}\ (U \in I_{2N+1})$

Proof. The result follows immediately from the fact that $\mathbb{P}(U \in I) = |I|$ for all intervals $I \subseteq [0,1]$, where $|I|$ is the length of interval I.

Definition 3.1. *Let ω_n stand for the sequence of random variables representing the overall amount of transcription lost after exactly n DNA damages have occurred. Note that if certain gene exhibits more than one damage this has no additional effect as we assume that a single damage is equally harmful and that transcription from that gene is permanently lost.*

Proposition 2. *Assume the above set-up. Then the expected transcription lost after n DNA damages reads*

$$\mathbb{E}(\omega_n) = \sum_{j=1}^{2N} \alpha_j (1 - (1-q_j)^n)$$

Proof. Let $\xi_k := \sum_{j=1}^{2N} \alpha_j \mathbb{1}\{U_k \in I_j, U_{k-1} \notin I_j, \ldots, U_1 \notin I_j\}$, where $(U_n)_{n\in\mathbb{N}}$ is a sequence of independent and identically distributed uniform $[0,1]$ random variables. Observe that Proposition 1 tells us that the entire path of damages can be captured by the sequence $(U_n)_{n\in\mathbb{N}}$. Moreover, we know that the additional transcriptions is lost if and only if a gene which was not previously hit exhibits a damage. Putting these two pieces of information together one can see that ξ_k precisely stands for the amount of additional transcription lost on the k^{th} DNA damage. Then clearly:

$$\omega_n = \sum_{k=1}^{n} \sum_{j=1}^{2N} \alpha_j \mathbb{1}\{U_k \in I_j, U_{k-1} \notin I_j, \ldots, U_1 \notin I_j\}$$

Using the fact that, for all measurable sets, $\mathbb{E}\mathbb{1}(A) = \mathbb{P}(A)$ we get:

$$\begin{aligned} \mathbb{E}(\omega_n) = \sum_{k=1}^{n} \sum_{j=1}^{2N} \alpha_j \mathbb{P}\{U_k \in I_j, U_{k-1} \notin I_j, \ldots, U_1 \notin I_j\} = \\ = \sum_{k=1}^{n} \sum_{j=1}^{2N} \alpha_j q_j (1-q_j)^{k-1} = \\ = \sum_{j=1}^{2N} \alpha_j q_j \sum_{k=1}^{n} (1-q_j)^{k-1} \end{aligned} \tag{3.3}$$

Hence the result follows from the independence of the sequence $(U_n)_{n\in\mathbb{N}}$ and interchanging of the order of summation.

Somewhat more involved calculation yields the behavior of the second moment. More specifically, we have the following result:

Theorem 3.1. *There exists a real number $\tau \in (0,1)$ and an absolute constant C such that $\mathbb{V}ar(\omega_n) \leq C\tau^n$ for all n simultaneously.*

Proof. We begin by computing

$$\begin{aligned}\left(\mathbb{E}\omega_n\right)^2 = \left(\sum_{k=1}^{2N} \alpha_j(1-(1-q_j)^n\right)^2 = \sum_{k=1}^{2N}\sum_{l=1}^{2N} \alpha_k\alpha_l(1-(1-q_k)^n(1-(1-q_l)^n \\ = \sum_{k=1}^{2N}\sum_{l=1}^{2N} \alpha_k\alpha_l - \sum_{k=1}^{2N}\sum_{l=1}^{2N} \alpha_k\alpha_l(1-q_k)^n - \sum_{k=1}^{2N}\sum_{l=1}^{2N} \alpha_k\alpha_l(1-q_l)^n + \sum_{k=1}^{2N}\sum_{l=1}^{2N} \alpha_k\alpha_l(1-q_k)^n(1-q_l)^n \\ = 1 - 2\sum_{k=1}^{2N} \alpha_k(1-q_k)^n + \sum_{k=1}^{2N}\sum_{l=1}^{2N}(1-q_k-q_l+q_kq_l)^n\end{aligned} \tag{3.4}$$

Moreover, we have:

$$\omega_n^2 = \sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{j=1}^{2N}\sum_{m=1}^{2N} \mathbb{1}\{U_k \in I_j, U_{k-1} \notin I_j, \ldots, U_1 \notin I_j, U_l \in I_m, U_{l-1} \notin I_m, \ldots, U_1 \notin I_1\} \tag{3.5}$$

and hence

$$\mathbb{E}\omega_n^2 = \sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{j=1}^{2N}\sum_{m=1}^{2N} \mathbb{P}\{U_k \in I_j, U_{k-1} \notin I_j, \ldots, U_1 \notin I_j, U_l \in I_m, U_{l-1} \notin I_m, \ldots, U_1 \notin I_1\} \tag{3.6}$$

Suppose $k = l$ and $j \neq m$. Non-zero contribution of these terms would imply that $U_k \in I_j$ and $U_k \notin I_j$ simultaneously which is clearly impossible. In other words, if $k = l$ then only the only cross terms which contribute are those when $j = m$ as well. We split the sum repeated sum above into three sub-cases, namely $k = l$, $k > l$ and $k < l$. We shall deal with the first two in detail while the third one is easily computed by symmetry. To this end we have:

Case 1: If $k = l$ then $j = m$ and whence by using the independence and distributional equality of the U_j's together with interchanging the order of summation, we see that the corresponding cross terms reduce to:

$$\begin{aligned}\sum_{k=1}^{n}\sum_{j=1}^{2N} \alpha_j^2\mathbb{P}\{U_k \in I_j, U_{k-1} \notin I_j, \ldots, U_1 \notin I_j\} = \sum_{k=1}^{n}\sum_{j=1}^{2N} \alpha_j^2 q_j(1-q_j)^{k-1} = \\ = \sum_{j=1}^{2N} \alpha_j^2(1-(1-q_j)^n)\end{aligned} \tag{3.7}$$

Case 2: $k > l$. The corresponding cross terms read:

$$\sum_{k=1}^{n}\sum_{l=1}^{k-1}\sum_{j=1}^{2N}\sum_{m=1}^{2N}\alpha_j\alpha_m\mathbb{P}\{U_k\in I_j,\ldots,U_{l+1}\notin I_j,U_l\notin I_j,U_l\in I_m,\ldots,U_1\notin I_m,U_1\notin I_j\} \tag{3.8}$$

Notice that if $j = m$ implies that $U_l \notin I_j$ and $U_l \in I_j$ and thus those terms with $j \neq m$ will yield zero contribution. Moreover, observe that $\{U_l \in I_m\}$ implies that $\{U_l \notin I_j\}$ since these intervals are disjoint by the very construction. In other words, $\{U_l \in I_m\} \subseteq \{U_l \notin I_j\}$ for all $j \neq m$. This implies that (1.8) simplifies to:

$$\sum_{k=1}^{n}\sum_{l=1}^{k-1}\sum_{j=1}^{2N}\sum_{m\neq j}\alpha_j\alpha_m q_j q_m(1-q_j)^{k-l-1}(1-q_j-q_m)^{l-1} \tag{3.9}$$

Furthermore, simple algebra and interchange of the order of summation also yields:

$$\begin{aligned}
&\sum_{k=1}^{n}\sum_{l=1}^{k-1}\sum_{j=1}^{2N}\sum_{m\neq j}\alpha_j\alpha_m q_j q_m(1-q_j)^{k-l-1}(1-q_j-q_m)^{l-1} = \\
&= \sum_{k=1}^{n}\sum_{j=1}^{2N}\sum_{m\neq j}\alpha_j\alpha_m q_j q_m(1-q_j)^{k}\sum_{l=1}^{k-1}(1-q_j-q_m)^{l-1}(1-q_j)^{-(l+1)} = \\
&= \sum_{k=1}^{n}\sum_{j=1}^{2N}\sum_{m\neq j}\alpha_j\alpha_m q_j q_m(1-q_j)^{k}\sum_{l=1}^{k-1}\Big(\frac{1-q_j-q_m}{1-q_j}\Big)^{l-1}(1-q_j)^{-2} = \\
&= \sum_{k=1}^{n}\sum_{j=1}^{2N}\sum_{m\neq j}\alpha_j\alpha_m q_j q_m(1-q_j)^{k-1}\Big(1-\Big(\frac{1-q_j-q_m}{1-q_j}\Big)^{k-1}\Big) = \\
&= \sum_{k=1}^{n}\sum_{j=1}^{2N}\sum_{m\neq j}\alpha_j\alpha_m q_j q_m(1-q_j)^{k-1} - \sum_{k=1}^{n}\sum_{j=1}^{2N}\sum_{m\neq j}\alpha_j\alpha_m q_j q_m(1-q_j-q_m)^{k-1} = \\
&= \sum_{j=1}^{2N}\sum_{m\neq j}\alpha_j\alpha_m(1-(1-q_j^n) - \sum_{j=1}^{2N}\sum_{m\neq j}\alpha_j\alpha_m\frac{q_j}{q_j+q_m}(1-(1-q_j-q_m)^n) = \\
&= \sum_{j=1}^{2N}\sum_{m=1}^{2N}\alpha_j\alpha_m(1-(1-q_j)^n) - \sum_{j=1}^{2N}\sum_{m=1}^{2N}\alpha_j\alpha_m\frac{q_j}{q_j+q_m}(1-(1-q_j-q_m)^n)
\end{aligned} \tag{3.10}$$

Finally, we shall put this expression in the form when the summations index j is unrestricted and hence the expression (3.10) further simplifies to:

$$\begin{aligned}
&\sum_{j=1}^{2N}\sum_{m=1}^{2N}\alpha_j\alpha_m(1-(1-q_j)^n)-\sum_{j=1}^{2N}\alpha_j^2(1-(1-q_j)^n)-\\
-\sum_{j=1}^{2N}\sum_{m=1}^{2N}\alpha_j\alpha_m\frac{q_j}{q_j+q_m}(1-(1-q_j-q_m)^n)+\sum_{j=1}^{2N}\frac{\alpha_j^2}{2}(1-(1-2q_j)^n)=\\
=\sum_{j=1}^{2N}\sum\alpha_j(1-(1-q_j)^n)-\sum_{j=1}^{2N}\alpha_j^2(1-(1-q_j)^n)-\\
-\sum_{j=1}^{2N}\sum_{m=1}^{2N}\alpha_j\alpha_m\frac{q_j}{q_j+q_m}(1-(1-q_j-q_m)^n)+\sum_{j=1}^{2N}\frac{\alpha_j^2}{2}(1-(1-2q_j)^n)
\end{aligned}\tag{3.11}$$

By symmetry, the cross-terms corresponding to the those indices such that $l \geq k$ read:

$$\begin{aligned}
&\sum_{j=1}^{2N}\sum\alpha_j(1-(1-q_j)^n)-\sum_{j=1}^{2N}\alpha_j^2(1-(1-q_j)^n)-\\
-\sum_{j=1}^{2N}\sum_{m=1}^{2N}\alpha_j\alpha_m\frac{q_j}{q_j+q_m}(1-(1-q_j-q_m)^n)+\sum_{j=1}^{2N}\frac{\alpha_j^2}{2}(1-(1-2q_j)^n)
\end{aligned}\tag{3.12}$$

Using suffix notation and summation convention we finalize the expression for the second moment of ω_n. Indeed, we have:

$$\begin{aligned}
\mathbb{E}\omega_n^2=2\sum_{j=1}^{2N}\alpha_j(1-(1-q_j)^n)-\sum_{j=1}^{2N}\alpha_j^2(1-(1-q_j)^n)-\\
-\sum_{j=1}^{2N}\sum_{m=1}^{2N}\alpha_j\alpha_m(1-(1-q_j-q_m)^n)+\sum_{j=1}^{2N}\alpha_j^2(1-(1-2q_j)^n)
\end{aligned}\tag{3.13}$$

Further simple algebra yields the expression for the variance of ω_n. Indeed, we have:

$$\begin{aligned}
\mathbb{V}ar\omega_n=\mathbb{E}\omega_n^2-(\mathbb{E}\omega_n)^2=2\sum_{j=1}^{2N}\alpha_j(1-(1-q_j)^n)-\sum_{j=1}^{2N}\alpha_j^2(1-(1-q_j)^n)-\\
-\sum_{j=1}^{2N}\sum_{m=1}^{2N}\alpha_j\alpha_m(1-(1-q_j-q_m)^n)+\sum_{j=1}^{2N}\alpha_j^2(1-(1-2q_j)^n)-\\
-1+2\sum_{j=1}^{2N}\alpha_j(1-q_j)^n-\sum_{j=1}^{2N}\sum_{m=1}^{2N}\alpha_j\alpha_m(1-q_j-q_m+q_jq_m)^n=\\
=2\sum_{j=1}^{2N}\alpha_j^2\big[(1-q_j)^n-(1-2q_j)^n)\big]+\sum_{j=1}^{2N}\sum_{m=1}^{2N}\alpha_j\alpha_m\big[(1-q_j-q_m)^n-(1-q_j-q_m+q_jq_m)^n\big]
\end{aligned}\tag{3.14}$$

Moreover let us define:

$$q^* := \min_{0 \leq j \leq 2N} q_j \tag{3.15}$$

Several applications of triangle inequality finally yield:

$$\mathbb{V}ar\omega_n \leq 3(1-q^*)^n \sum_{j=1}^{2N} \alpha_j^2 \leq 3(1-q^*)^n \tag{3.16}$$

and hence the proof is complete.

As damages accumulate the overall amount of transcription lost increases and eventually approaches 1 (100%). We have seen that $\lim_{n\to\infty} \mathbb{V}ar(\omega_n) = 0$ and $\lim_{n\to\infty} \mathbb{E}(\omega_n) = 1$ and hence the model is in line with these simplistic demands. However, much more is true but further definitions and results are needed. To this end we have:

Definition 3.2. *A sequence of random variables $(X_n)_{n\in\mathbb{N}}$ is said to converge in L_p to some random variable X, written $X_n \xrightarrow{L_p} X$, if and only if $\lim_{n\to\infty} \mathbb{E}(|X_n - X|^p = 0$*

Definition 3.3. *A sequence of L_1 random variables $(X_n)_{n\in\mathbb{N}}$ is said to be uniformly integrable if and only if*

$$\lim_{a\to\infty} \sup_{n\in\mathbb{N}} \mathbb{E}(|X_n| \mathbb{1}\{|X_n| \geq a\} = 0$$

Definition 3.4. *We say that a sequence of random variables $(X_n)_{n\in\mathbb{N}}$ converges to a random variable X in probability, written $X_n \xrightarrow{\mathbb{P}} X$ if and only if, for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$*

Definition 3.5. *A sequence of random variables $(X_n)_{n\in\mathbb{N}}$ converges almost surely to a random variable X, written $X_n \xrightarrow{a.s.} X$ if and only if there exists some measurable set A with $\mathbb{P}(A) = 1$ and $X_n(\omega) \to X(\omega)$ for all $\omega \in A$.*

Proposition 3. *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables and suppose there exists a dominating L_1 random variable Y with $|X_n| \leq Y$ for all $n \in \mathbb{N}$. Then the family $(X_n)_{n\in\mathbb{N}}$ is uniformly integrable.*

Proof. Please see page 183 of ([22]) for details.

Proposition 4. *Let $(X_n)_{n\in\mathbb{N}}$ be a monotone sequence of random variables that converges in probability to some random variable X. Then $(X_n)_{n\in\mathbb{N}}$ converges to X almost surely.*

Proof. Please see page 195 of ([22]) for details.

Theorem 3.2. *Suppose $p \geq 1$ and $(X_n)_{n\in\mathbb{N}} \in L^p$. The following statements are equivalent*

(a) $(X_n)_{n\in\mathbb{N}}$ is L_p convergent.

(b) $|X_n|^p_{n\in\mathbb{N}}$ is uniformly integrable and $(X_n)_{n\in\mathbb{N}}$ converges in probability.

Proof. Please see page 194 of ([22]) for details.

Theorem 3.3. *The following statements hold true:*

(a) $\omega_n \overset{a.s.}{\to} 1$

(b) $\omega_n \overset{L_p}{\to} 1$ *for all* $p \geq 1$

Proof.

By definition $\omega_n \leq 1$ and whence

$$\mathbb{E}|\omega_n - 1| = 1 - \mathbb{E}\omega_n = 1 - \sum_{j=1}^{2N} \alpha_j (1 - (1 - q_j)^n) \to 0 \tag{3.17}$$

as $n \to \infty$ since all genes have strictly positive lengths. In other words, $\omega_n \overset{L_1}{\to} 1$. Then by Theorem 1.2 $\omega_n \overset{\mathbb{P}}{\to} 1$. However, as damages can only accumulate in time, $\omega_n \geq \omega_{n-1}$ for all $n \in \mathbb{N}$ and hence $(\omega_n)_{n\in\mathbb{N}}$ is a monotonically increasing. Thus, by Proposition 4, $\omega_n \overset{a.s.}{\to} 1$. Moreover, as $|\omega_n| \leq 1$ we have $(\omega_n)_{n\in\mathbb{N}} \in L^p$ for all $p \in \mathbb{R}^+$. Moreover by Theorem 1.2 $(|\omega_n|^p)_{n\in\mathbb{N}}$ is uniformly integrable and the proof is complete.

We are also interested in studying the least amount of damages needed to block some prescribed level of transcription, β, say. More specifically, the levels $\beta \in \{0.3, 0.5, 0.7, 0.9\}$ are of biological interest. Mathematically, this is captured by studying the random variable T which is the first hitting time of the sequence $(\omega_n)_{n\in\mathbb{N}}$ to level β. More specifically:

$$T = \inf\{n \geq 0 : \omega_n \geq \beta\} \tag{3.18}$$

We now provide a closed form expression for the distribution of T. By using the law of total probability we have:

$$\mathbb{P}(T=n)=\mathbb{P}(\omega_n\geq\beta,\omega_{n-1}<\beta)=\mathbb{P}(\xi_1+\xi_2+\cdots+\xi_n\geq\beta,\xi_1+\xi_2+\cdots+\xi_{n-1}<\beta)=$$
$$=\sum_{i_1=1}^{2N+1}\sum_{i_2=1}^{2N+1}\cdots\sum_{i_{n-1}=1}^{2N+1}\mathbb{P}(\xi_1+\xi_2+\cdots+\xi_n\geq\beta,\xi_1+\xi_2+\cdots+\xi_{n-1}<\beta|U_1\in I_{i_1},\ldots,U_{n-1}\in I_{i_{n-1}})=$$
$$=\sum_{i_1=1}^{2N+1}\sum_{i_2=1}^{2N+1}\cdots\sum_{i_{n-1}=1}^{2N+1}q_{i_1}\ldots q_{i_n}\sum_{l^*:}q_{l^*} \tag{3.19}$$

where l^* corresponds to all those indices satisfying the following relation:

$$l^*\notin\{i_1,\ldots,i_{n-1},2N+1\},$$
$$\alpha_{l^*}\geq\alpha_{i_1}\mathbb{1}\{i_1\neq 2N+1\}-\alpha_{i_2}\mathbb{1}\{i_2\neq i_1,i_2\neq 2N+1\}-\ldots\alpha_{i_{n-1}}\mathbb{1}\{i_{n-1}\notin\{i_1,\ldots,i_{n-2},2N+1\}\} \tag{3.20}$$

We are also interested in a problem of computing the probability distribution of the number of damages needed to "switch off" both copies of some specific gene, that is number of damages needed to see the transcription associated with this gene falling to 0. To be more specific, we have:

Proposition 5. *Let Ω^i stand for the random variable representing the smallest number of damages needed for both copies of gene i (in further text these will be labeled i and i^*) to exhibit a damage, that is, the first moment in time when the transcription associated with this gene falls permanently to 0. Then:*

$$\mathbb{P}(\Omega^i=n)=2q_i((1-q_i)^{n-1}-(1-2q_i)^{n-1})$$

Proof.

$$\mathbb{P}(\Omega^i=n)=\mathbb{P}(\Omega^i=n\text{ and }i\text{ is hit last})+\mathbb{P}(\Omega^i=n\text{ and }i^*\text{ is hit last})=$$
$$2\mathbb{P}(\Omega^i=n\text{ and }i\text{ is hit last}) \tag{3.21}$$

as i and i^* are equally likely to exhibit a DNA damage. Observe that the event $\{\Omega^i=n\text{ and }i\text{ is hit last}\}$ can only occur if and only if among the first $n-1$ DNA damages at least one occurs in gene i^* while gene i is hit for the first time precisely on the n^{th} DNA damage. By conditioning on the number of i^*s among the first $n-1$ DNA damages we have:

$$\mathbb{P}(\Omega^i=n\text{ and }i\text{ is hit last})=q_i\sum_{k=1}^{n-1}\binom{n-1}{k}q_i^k(1-2q_i)^k=$$
$$=q_i\left(\sum_{k=0}^{n-1}\binom{n-1}{k}q_i^k(1-2q_i)^k-(1-2q_i)^{n-1}\right)=q_i((1-q_i)^{n-1}-(1-2q_i)^{n-1}) \tag{3.22}$$

and the proof is complete.

4 Simulation study

Although the exact expression for the probability distribution of the random variable T is available, its practical value is rather questionable due to the apparent computational complexity needed to implement the closed form solution. However, we provide a simple algorithm which yields the approximate distribution of T in a computationally feasible manner. Indeed, we simulate the process $(\omega_n)_{n\in\mathbb{N}}$ specified number of times and record the value of T on each such run to obtain the corresponding histograms. The pseudo-algorithm is presented below:

$\omega = 0$; $T = 0$;
Generate a uniform [0,1] random variable U_1. Find n such that $U_1 \in I_n$ for $n \in \{1, 2, \ldots, 2N + 1\}$
If $n = 2N + 1$ $T++$;
else $\omega = \omega + \alpha_n$
If $\omega \geq \beta$, Stop and return $T = 1$ else generate a random variable U_2, U_2 independent of U_1 and uniformly distributed on [0,1]
Find m such that $U_2 \in I_m$
if ($m = 2N + 1$ or $U_1 \in I_m$) T++ ;
else $\omega = \omega + \alpha_m$
if $\omega \geq \beta$ T=2;
otherwise continue generating new independent and identically distributed uniform [0,1] random variables until you eventually reach the prescribed level β
return T;

Transcription loss is computed based on relative levels of nascent RNA transcription for each gene from three biological replicates. Data set spans 1331939K total base pairs, 9661 Genes, including both alleles of each gene. The instances of DNA damage were inflicted uniformly at random throughout the genome until a specified loss of transcription has been reached. This procedure was repeated 100 times to generate a representative histogram of the first hitting time T.

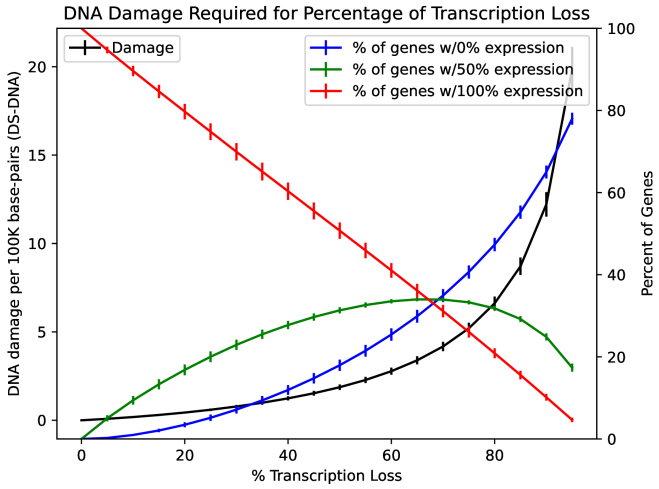

Figure 1 Relationship between the percentage of fully transcribed genes (blue), genes with 50 percent transcription loss (green) and genes with completely blocked transcription (red) and the number of DNA damages

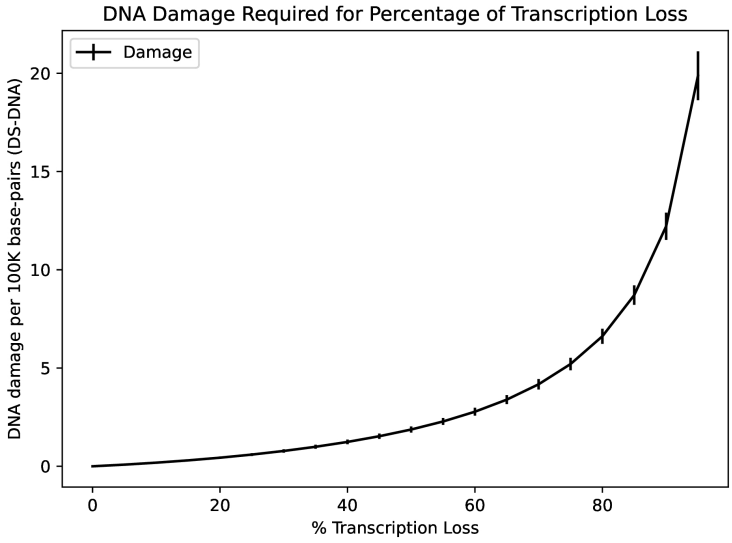

Figure 2 Relationship between transcription loss and a number of accumulated DNA damages

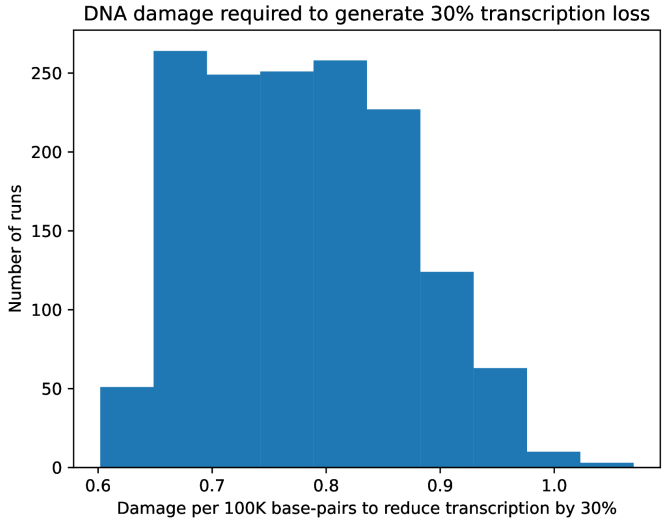

Figure 3 Histogram of the probability distribution of the first hitting time for the number of accumulated DNA damages resulting in 30 percent transcription loss

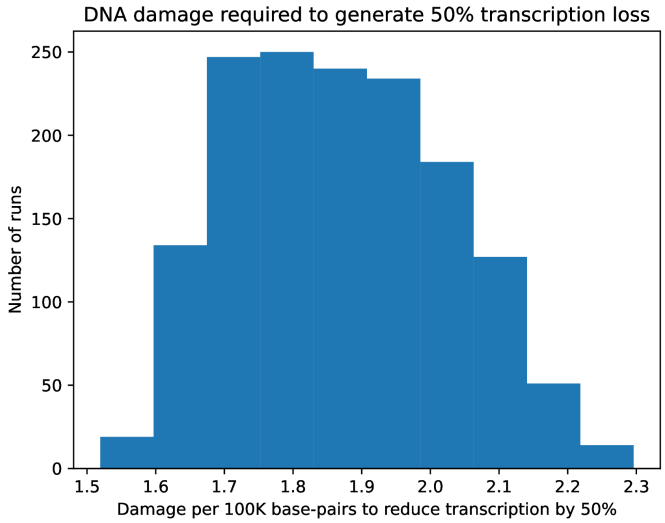

Figure 4 Histogram of the probability distribution of the first hitting time for the number of accumulated DNA damages resulting in 50 percent transcription loss

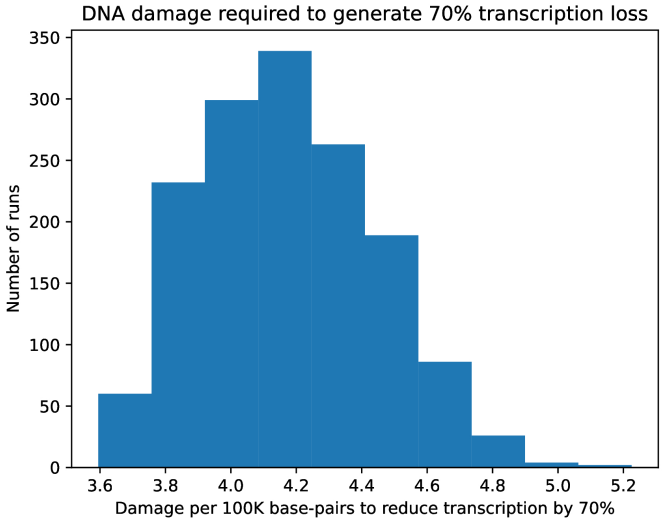

Figure 5 Histogram of the probability distribution of the first hitting time for the number of accumulated DNA damages resulting in 70 percent transcription loss

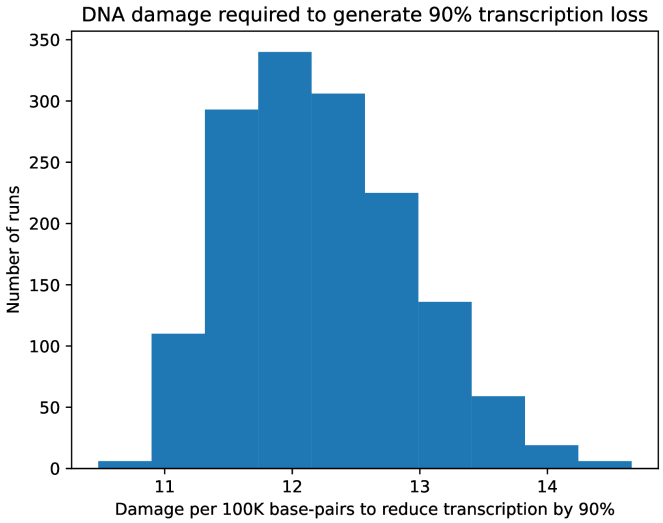

Figure 6 Histogram of the probability distribution of the first hitting time for the number of accumulated DNA damages needed for 90 percent transcription loss

5 Conclusion

We develop a simple mathematical model of DNA transcription loss due to the accumulated DNA damages. More specifically, we provide closed form formulae for the first two moments of the distribution of the transcription lost upon specified number of DNA damages. The associated stochastic process is demonstrated to converge to 1 as the number of damages tends to infinity for a variety of probabilistic convergence modes, including almost sure convergence and convergence in L^p, for all $p \geq 1$. Moreover, we provide closed form formulae for the probability distribution function for the random variable representing the number of damages needed to switch off both copies of a gene. Furthermore, the closed form formula for the distribution of the first hitting time of specified level of blocked transcription is provided. Unfortunately, direct application of this formula is practically infeasible due to its computational complexity, however, we have implemented a simple algorithm in a simulation study to draw statistical inference on this biologically important quantity. We plan to subsequently generalize this model further accounting for the DNA repair mechanism and study the effect on protein synthesis. Finally, we will use analytic and computational inference in conjunction with experimental in vivo and in vitro data to advance our understanding of the implications of transcription loss in aging.

References

[1] Alexander, P. The role of DNA lesions in the processes leading to aging in mice, *Symp Soc Exp Biol* (**1967**)

[2] Gensler, H. L. and Bernstein, H. DNA damage as the primary cause of aging, *Q Rev Biol* (**1981**)

[3] Bernstein, C. and Bernstein, H Aging, Sex, and DNA Repair, *San Diego: Academic Press* (**1991**)

[4] Ames, B. N and Gold, L. S. Endogenous mutagens and the causes of aging and cancer, *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* (**1991**)

[5] Holmes, G. E., Bernstein, C. and Bernstein, H. The role of DNA lesions in the processes leading to aging in mice, *Oxidative and other DNA damages as the basis of aging: a review* (**1992**)

[6] Rao, K. S., Loeb, L. A. DNA damage and repair in brain: relationship to aging, *Mutation Research/DNAging* (**1992**)

[7] Ames, B. N., Shigenaga, M. K. and Hagen, T. M. Oxidants, antioxidants, and the degenerative diseases of aging *Proceedings of the National Academy of Sciences* (**1993**)

[8] Lu, T, Pan, Y, Kao, SY, Li, C, Kohane, I, Chan, J and Yankner, BA Gene regulation and DNA damage in the ageing human brain, *Nature* (**2004**)

[9] Bernstein H, Payne CM, Bernstein C, Garewal H and Dvorak K Cancer and aging as consequences of un-repaired DNA damage, *) Nova Science Publishers, Inc., New York,* (**2008**)

[10] Hamilton, M. L., Van Remmen, H., Drake, J. A., Yang, H., Guo, Z. M., Kewitt, K., Walter, C. A. and Richardson, A. Does oxidative damage to DNA increase with age?, *Proceedings of the National Academy of Sciences of the United States of America* (**2001**)

[11] Mecocci, P., FanĂŁ, G., Fulle, S., MacGarvey, U., Shinobu, L., Polidori, M. C., Cherubini, A, Vecchiet, J., Senin, U. and Beal, M. F. Age-dependent increases in oxidative damage to DNA, lipids, and proteins in human skeletal muscle , *Free Radic Biol Med.* (**1999**)

[12] Helbock, HJ, Beckman, KB and Shigenaga, MK DNA oxidation matters: the HPLC-electrochemical detection assay of 8-oxo-deoxyguanosine and 8-oxoguanine, *Proc. Natl. Acad. Sci. U.S.A.* (**1998**)

[13] Hashimoto, K, Takasaki, W, Sato, I and Tsuda, S DNA damage measured by comet assay and 8-OH-dG formation related to blood chemical analyses in aged rats, *J Toxicol Sci.* (**2007**)

[14] Werner, O. On cataract in conjunction with scleroderma. Otto Werner, doctoral dissertation, *Advances in Experimental Medicine and Biology* (**1985**)

[15] Hutchinson J. Case of congenital absence of hair, with atrophic condition of the skin and its appendages, in a boy whose mother had been almost wholly bald from alopecia areata from the age of six, *Lancet* (**1886**)

[16] Neill CA and Dingwall MM. A Syndrome Resembling Progeria: A Review of Two Cases, *Archives of Disease in Childhood* (**1950**)

[17] Marcel O., Stephan H., Jan G.H., Deborah E.B., Tomas L and Bernd E, Age-related and tissue-specific accumulation of oxidative DNA base damage in 7,8 dihydro-8-oxoguanine-DNA glycosylse (Ogg1) deficient mice *Carcinogenesis* (**2001**)

[18] Ruijie S., Murat A. Stochastic modeling of aging cells reveals how damage accumulation, repair, and cell-division asymmetry affect clonal senescence and population fitness, *BMC Bioinformatics* (**2019**)

[19] Griffits A., Miller J., Suzuki D., Lewontin R. and Gelbart W.M. An Introduction to Genetic Analysis 7th edition, *New York W.H. Freeman* (**2000**)

[20] Jinchuan H, Jason D. L., Aziz S., and Sheera A Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution, *PNAS* (**2016**)

[21] Haitham S., Rajendra K., Jacob L., Ludvig L and Per S. Highly interacting regions of the human genome are enriched with enhancers and bound by DNA repair proteins, (**1967**)

[22] Sidney I. Resnick A Probability Path, *Birkhauser* (**1999**)

[23] Gyenis, A., Chang, J., Demmers, J.J.P.G. et al. Genome-wide RNA polymerase stalling shapes the transcriptome during ageing, *Nature Genetics*, 55, 268-279, (**2023**)

[24] Koehler, K., Ferreira, P., Pfander, B. and Boos, D. The Initiation of DNA Replication in Eukaryotes, *Springer, Cham*, 443-460, (**2016**)

[25] Iyama, T., Wilson, D.M. DNA repair mechanisms in dividing and non-dividing cells, *DNA Repair*, 620-636, (**2013**)