3 tera-basepairs as a fundamental limit for robust DNA replication

Al Mamun, M.; Albergante, L.; Blow, J. J.; Newman, T. J.

Published in:
Physical Biology

DOI:
10.1088/1478-3975/ab8c2f

Publication date:
2020

Physical Biology

PAPER • **OPEN ACCESS**

3 tera-basepairs as a fundamental limit for robust DNA replication

To cite this article: M Al Mamun *et al* 2020 *Phys. Biol.* **17** 046002

Physical Biology

OPEN ACCESS

PAPER

3 tera-basepairs as a fundamental limit for robust DNA replication

M Al Mamun[1,2,6], L Albergante[1,3,4], J J Blow[1] and T J Newman[1,5,6]

[1] School of Life Sciences, University of Dundee, Dundee DD1 5EH, United Kingdom
[2] CIB-CSIC, Madrid 28040, Spain
[3] U900, Institut Curie, Paris 75005, France
[4] Sensyne Health, Oxford OX4 4GE, United Kingdom
[5] Solaravus, Cupar, Fife, United Kingdom
[6] Author to whom any correspondence should be addressed.

E-mail: mohammedal.mamun@cib.csic.es and tjnewman@solaravus.com

Keywords: DNA replication, embryo development, polyploidy, robustness, eutely, double fork stalls, theoretical analysis

RECEIVED
9 December 2019

REVISED
31 March 2020

ACCEPTED FOR PUBLICATION
22 April 2020

PUBLISHED
30 June 2020

Abstract

In order to maintain functional robustness and species integrity, organisms must ensure high fidelity of the genome duplication process. This is particularly true during early development, where cell division is often occurring both rapidly and coherently. By studying the extreme limits of suppressing DNA replication failure due to double fork stall errors, we uncover a fundamental constant that describes a trade-off between genome size and architectural complexity of the developing organism. This constant has the approximate value $N_U \approx 3 \times 10^{12}$ basepairs, and depends only on two highly conserved molecular properties of DNA biology. We show that our theory is successful in interpreting a diverse range of data across the Eukaryota.

1. Introduction

Organisms are made from cells, and their functional and morphological integrity relies upon the integrity of cellular processes, particularly cell division. In turn, this relies upon the integrity of the molecular process of DNA replication [1]. Thus, there is a direct link across multiple biological scales, connecting organismal robustness to genomic fidelity. Indeed, it is vital for developmental and other growth processes in organisms that the DNA in each new cell is as faithful as possible to the original zygotic genome. Errors in DNA replication will inevitably occur and cells have sophisticated means to identify and repair such errors. However, repairing DNA errors, particularly gross ones, is time-consuming, and such a bottleneck in a given cell could interfere badly with higher-level coordinated cell division processes. This is particularly relevant in embryo development, which for many organisms is highly streamlined, with 'stripped-down' cell division cycles (e.g. cleavage divisions) operating across the embryo in synchrony [2]. The coherent generation of significant numbers of correctly differentiated cells enables the formation of complex architectures that constitute the emerging morphology of the organism. For many organisms development must be rapid to allow the nascent life form to function as an autonomous agent, able to compete for resources and evade predation in a hostile environment.

Thus, a tension exists between the robustness and the rapidity of development; between the requirements of integrity of DNA replication during cell division and of the speedy emergence of autonomously functional biological form. This can be restated more concisely as a tension between information fidelity and organismal functionality. We investigate this by considering an important example of DNA replication error for which repair is possible but costly in time, namely, double fork stalls (DFS) [3, 4]. We shall be able to quantify in a surprisingly simple way the tension described above, which, in a developmental context, takes the form of a trade-off between genome size (information complexity) and embryonic cell number (architectural complexity). This trade-off is expressed in terms of a single constant which we denote by N_U, and which has dimensions of DNA length. We believe N_U to be highly conserved across the eukaryotes. It has the approximate value 3 Tbp, i.e. $N_U \approx 3 \times 10^{12}$ bp.

The outline of this paper is as follows. We provide a short overview of the biology of DFS and summarise a recent theory that has successfully captured much of the experimental data for DFS in both yeast cells and human cell lines. We use one element of this theory to derive the main result of this paper, and then proceed to test this against data from a diverse range of biological examples drawn from the Eukaryota, including eutely, syncytial development and polyploidy. We end with a summary of our results and a discussion of extensions of our theory. A guide to notation and further calculational details are provided in the appendix.

2. Background to DFS and a recent quantitative theory

Replication of DNA is initiated at multiple sites, called replication origins (ROs), situated along the DNA chain. In order to prevent any RO from firing twice in the same cell cycle (which would cause sections of DNA to be replicated twice in the same cell cycle), eukaryotic cells divide the process of replication into two non-overlapping phases [5]. From late mitosis until the end of G1, before DNA synthesis begins, cells 'license' ROs for use by loading them with double hexamers of the MCM2-7 (minichromosome maintenance) proteins. Once cells enter S phase, when RO firing can occur, no further ROs can be licensed. When an RO is activated ('fires') during S phase of the cell cycle, two replication forks proceed with replication in opposite directions along the DNA, each driven by one of the two MCM2-7 hexamers loaded onto the origin. Note that only a subset of licensed ROs fire during any particular S phase, with the remaining 'dormant' origins remaining as potential backups for use if problems occur to the active replication forks [3, 4]. If a replication fork encounters a dormant ('unfired') RO, replication continues past the dormant origin and the MCM2-7 loaded onto it is removed (the dormant origin becomes 'unlicensed'). This prevents re-replication of already replicated DNA [5]. The complex of proteins at a given replication fork is called a 'replisome' and consists of an assembly of molecular machines working in a coordinated fashion to replicate the DNA rapidly (ca 50 bp s^{-1} in eukaryotes) and accurately (ca single nucleotide error rate of 10^{-9}) [1]. Despite this sophistication, replication forks can fail through rare irreversible stalling. This is typically not problematic, as the unreplicated DNA lying ahead of the stalled fork will eventually be replicated by another fork moving in the opposite direction having been initiated by an RO upstream of the stalling event. Very rarely though a severe error can occur, a DFS. In this situation, two converging replication forks irreversibly and independently stall with no dormant RO available in the stretch of unreplicated DNA lying between them. A more detailed description of DFS with schematic illustrations can be found in [3].

A simple theory of DFS statistics has recently been developed and is successful in predicting error rates and RO distributions for genomes spanning Mbp (e.g. yeast) to Gbp (e.g. human) [6, 7]. The theory has a single *a priori* unknown parameter q, the genome-wide average probability of a single fork stall per nucleotide replication. Fits of the theory to various experimental data have consistently indicated the approximate value $q \approx 5.8 \times 10^{-8}\,\mathrm{bp}^{-1}$. This parameter can be recast as the length of DNA replicated before a 50% chance of a single fork stall, which we denote by N_s, and which has the approximate value $N_s = \ln 2/q \approx 12\,\mathrm{Mbp}$. Henceforth we shall exclusively use the symbol N, with one of a number of subscripts, to denote various length scales of DNA that arise in the theory. A complete list of the symbols used is given in the appendix to aid the reader.

In previous applications of the theory, to yeast cells [6] and human cell lines [7, 8], a key experimental input was the set of inter-RO separations, which has a mean value typically of order 10 kbp in these examples. The theory was able to explain how this scale of RO separation leads to small tolerable DFS error rates in single cell divisions. The theory was also able to show that the observed RO distributions are optimized to constrain the number of DFS errors in a single cell division for the very different genome sizes under consideration.

Here, we consider a different situation; that of extreme elimination of DFS errors. We have foremost in our minds the case of rapid coordinated cell divisions, for instance in early embryo development, but our theory has wider applicability than this. Note, we are not concerned with the 'timing question' of ensuring complete DNA replication within a single cell in a preset time period, which has had considerable previous study using other theoretical approaches [9–11].

3. Derivation of the central result

This work was spurred by the experimental finding of very high levels of RO licensing proteins in the cells of the developing *Xenopus* embryo [12–15]. These studies suggest that the total amount of MCM2-7 in the *Xenopus* egg is sufficient to provide a double hexamer at least every 400 bp throughout the first 12 embryonic cell cycles until zygotic transcription starts (at the mid-blastula transition). Although the spacing between fired origins has been measured to be ∼10 kbp [12], the density of dormant origins is at least ten times higher than this [16, 17]. One can postulate that for an embryonic cell to absolutely minimise its chance of a DFS error, it would, prior to S phase, saturate its DNA with ROs. The finest scale at which this is possible is the 'quantum'

of eukaryotic DNA organisation, i.e. the nucleosome (and accompanying inter-nucleosome regions of DNA) [1]. The length of nucleosome linkers across eukaryotes ranges between ca 20–90 bp, and the footprint of licensing molecules is ca 60 bp [18–21]. Therefore an average inter-nucleosome distance of ∼60 bp allows for an essentially whole-genome saturation with ROs. For the purposes of our theory, we therefore consider the DNA as quantised on the periodic scale of nucleosomes and their accompanying inter-nucleosome regions, which we denote by N_n, and which has a value of ca 200 bp [1]. We define the parameter ρ to be the probability that a given inter-nucleosome region is occupied by an RO. In the limit of $\rho \to 1$ the DNA is saturated with ROs, the number of which across a genome of size N_g is in this case given by N_g/N_n.

In section B of the appendix we present the theory for the general case of $0 < \rho \leqslant 1$. For the main results of this paper we are interested in the extreme case of $\rho \to 1$, for which a short and straightforward derivation of the theory is possible, as we now describe.

A basic ingredient of the recent theory of DFS error rates is the probability of a DFS event in a region of DNA of size N. For $1 \ll N \ll N_s$ this has the form (see equations (A8)and (A16) in [6]):

$$P_{\mathrm{DFS}}(N) = \frac{1}{2}q^2N^2 = \frac{(\ln 2)^2}{2}\left(\frac{N}{N_s}\right)^2. \quad (1)$$

Thus, if we assume that every inter-nucleosome region is occupied by an RO, the probability of a DFS event within a 200 bp nucleosomal region N_n is

$$P_{\mathrm{DFS}}(N_n) = \frac{(\ln 2)^2}{2}\left(\frac{N_n}{N_s}\right)^2 \approx 6.6 \times 10^{-11}, \quad (2)$$

which is exceedingly small, as expected.

We now consider a total amount of DNA of length N_t to be replicated, all of which is saturated with ROs as described above. This total amount of DNA may reside inside a single cell or may be distributed among more than one cell, depending upon the application of interest. Given that potential DFS errors within each nucleosomal stretch of DNA are independent events, the probability of no DFS errors occurring within the entire replication process is given by $(1 - P_{\mathrm{DFS}}(N_n))$ raised to the power of N_t/N_n. Thus, the probability of one or more DFS errors occurring is

$$P_{\mathrm{error}}(N_t) = 1 - (1 - P_{\mathrm{DFS}}(N_n))^{N_t/N_n}. \quad (3)$$

Given the extremely small value of $P_{\mathrm{DFS}}(N_n)$ this expression may be rewritten as

$$P_{\mathrm{error}}(N_t) = 1 - \exp\left(-\frac{N_t}{N_n}P_{\mathrm{DFS}}(N_n)\right). \quad (4)$$

Now, focussing on the argument of the exponential, we have from equation (1):

$$\frac{N_t}{N_n}P_{\mathrm{DFS}}(N_n) = \frac{N_t}{N_n} \times \frac{1}{2}q^2{N_n}^2 = UN_t, \quad (5)$$

where we have introduced the fundamental constant

$$U = \frac{1}{2}q^2N_n \approx 3.3 \times 10^{-13}\,\mathrm{bp}^{-1}. \quad (6)$$

We describe U as 'fundamental' as it comprises two molecular constants which are strongly conserved across eukaryotic life: i) the per nucleotide spontaneous stalling probability of the DNA replication machinery and ii) the average periodicity of nucleosomes.

It is more convenient for our purposes to define the inverse of U, which has dimensions of DNA length. We define

$$N_U = 1/U \approx 3.0 \times 10^{12}\,\mathrm{bp}. \quad (7)$$

Given that N_U is simply the inverse of U, the adjective 'fundamental' applies equally well to it, and thus we posit that the value of three tera-basepairs (i.e. 3 Tbp) is a fundamental scale in rapid, large-scale DNA replication and the biology that depends upon it. Our results, presented shortly, appear to support this view.

Returning to our expression for $P_{\mathrm{error}}(N_t)$ in equation (4), and using equations (5)–(7), we have our central result:

$$P_{\mathrm{error}}(N_t) = 1 - \exp\left(-\frac{N_t}{N_U}\right). \quad (8)$$

If the total amount of DNA under consideration has length much less than N_U, i.e. much less than 3 Tbp, then the expression can be simplified to

$$P_{\mathrm{error}}(N_t) = \frac{N_t}{N_U} \ll 1. \quad (9)$$

In anticipation of the biological examples to follow, we can consider two general cases.

Case I: this occurs when the total amount of DNA to be replicated is distributed among more than one cell (each of which we assume to have the same genomic content). We define the genome size N_g of each cell to be the number of basepairs in a haploid set of chromosomes. If we assume these cells to be diploid, and for there to be a final count of M_c cells (starting from a single cell after $M_c - 1$ cell divisions), then the total amount of DNA to be replicated is $N_t = 2(M_c - 1)N_g$. If the cells saturate their DNA with ROs in order to ensure the smallest chance of DFS errors, then assuming that $P_{\mathrm{error}}(N_t)$ is small (and taking for simplicity $M_c \gg 1$) we have from our theory above

$$P_{\mathrm{error}}(N_t) = \frac{2M_cN_g}{N_U} \ll 1, \quad (10)$$

and consequently the inequality:

$$M_cN_g \ll N_U. \quad (11)$$

This expression encapsulates the trade-off between genome size and the number of cells involved in the coordinated cell division process. The product of the 'architectural complexity' (M_c) and the 'informational complexity' (N_g) are bounded by N_U; they cannot be simultaneously increased such that their product exceeds N_U without introducing costly forms of DNA error repair. Typically, we would imagine such a process occurring during embryonic development, though examples involving rapid, coordinated cell divisions in adult organisms could also be relevant.

Case II: this occurs when the entirety of the DNA to be replicated is within one cell. Defining M_p as the degree of polyploidy, we have $N_t = M_p N_g$. If the cell saturates its DNA with ROs in order to minimise the chance of DFS errors, then following the same line of argument as above we have the inequality:

$$M_p N_g \ll N_U, \tag{12}$$

which encapsulates a trade-off between genome size and degree of polyploidy for such cells.

Before turning to some biological examples, we briefly discuss the more general case of $\rho < 1$. As mentioned above, section B in the appendix provides a derivation of the central result (the analogue of equation (8)) for arbitrary values of ρ. This general result is analysed in section C of the appendix resulting in two useful observations. Firstly, equation (Aix) shows that as ρ decreases from unity, the DFS error rate increases dramatically as $2/\rho$. This will provide strong pressure to keep the RO density close to saturation ($\rho = 1$) when replication of DNA content close to 3 Tbp is required. Second, equation (Axi) provides a lower bound on ρ which can be calculated using knowledge only of the theoretical error rate at saturation (i.e. inserting N_t into equation (8)) and the experimentally observed failure rate of the biological process under consideration. This bound will prove useful when more detailed data becomes available of RO distributions during early embryonic processes. We will give an example of the use of this bound below, when discussing eutelic organisms.

4. Testing the central result using specific biological examples

In this section, we test our central results against experimental data. Relating to case I, we look at two examples: i) eutelic organisms from across the Eukaryota, and ii) the syncytial phase of *Drosophila* development. We then turn briefly to case II, using examples of high degree polyploidy from *Drosophila*, mouse and human cell types.

4.1. Eutely

We start by considering what is perhaps the most highly coordinated mode of development, in which the form of the organism emerges from a completely prescribed set of cell divisions, such that the number of cells and their individual differentiated states are precisely defined at each stage of development. This process is called eutely and has been adopted across diverse branches of the eukaryotes [22]. The eutelic organism has a predictable number of cells and after cell division ceases it grows larger through each cell increasing in size. In terms of our theory, we would expect the inequality in equation (11), namely $M_c N_g \ll N_U$, to place a profound constraint, simultaneously, on the genome size and cell number of eutelic organisms. This is under the assumption, of course, that the cell divisions during development are highly coordinated and rapid such that significant time spent repairing gross errors from DFS is not possible.

To test this idea, we examine eutelic organisms for which cell number and genome size are known and then compare their product to the fundamental constant N_U. A more precise test is also possible, since use of equation (8) with the ratio of $2M_c N_g$ to N_U substituted in the argument of the exponential gives the probability of one or more DFS errors. If such errors are essentially lethal for eutelic embryos, then this ratio provides an estimate (more precisely, a lower bound) for the failure rate of development of such embryos.

Table 1 provides data for three species of eutelic organisms which sit within three distinct branches of the eukaryotes: the nematode *Caenorhabditis elegans* [1, 23], the tardigrade *Hypsibius dujardini* [24, 25] and the rotifer *Brachionus calyciflorus* [26, 27]. We note a remarkable similarity of the cell number counts and genomic complexity of the organisms despite their very distinct taxonomies, morphologies and environments. The data is in good accord with the predictions of our theory. Our estimates of DFS errors, assuming saturation of the DNA with ROs, are also consistent in being slightly smaller than the observed developmental failure rates of the organisms (denoted by P_{obs}). This does not constitute conclusive proof that DFS errors ultimately limit the complexity of eutelic organisms. Experiments are required to demonstrate this; for instance, to show that the DNA of cells in eutelic development are saturated with ROs, or to show that those embryos that fail contain cells that are unable to complete timely divisions due to the occurrence of one or more DFS errors. We can also use equation (Axi) to estimate lower bounds of ρ (denoted by ρ_{min}) using the error rates in columns 5 and 6 (mean value) of the table. These bounds are provided in column 7, and range from 0.67 to 0.89 indicating that all organisms are utilising near saturation in order to control DFS errors. In the pre-gastrula *C. elegans* embryo, the number of identified ROs is ∼15 000 (although noting that large parts of the genome in the microarray-based study were missed due to the technical limitations in accessing highly repetitive

Table 1. Data and theory predictions for three eutelic organisms.[a]

Species	N_g(Mbp)	M_c	M_cN_g(Gbp)	P_{error}	P_{obs}	ρ_{min}
C. elegans	≈100	≈1000	≈100	≈6%	11–12%	0.67
H. dujardini	≈100	≈1000	≈100	≈6%	7–9%	0.86
B. calyciflorus	≈65	≈1000	≈65	≈4%	<5%	0.89

[a] Note, P_{error} is calculated from equation (8) and ρ_{min} from equation (Axi).

sequences) [28]. If the abundance of dormant origins in this organism is as high as in *Xenopus* (ten times that of active ROs), then our calculated ρ_{min} of 0.67 suggests around 30 000 active ROs genomewide. This rough estimate is twice that observed in the microarray experiment possibly suggesting that half the origins licensed are in highly repetitive regions of the genome.

4.2. Syncytial development

Our analysis indicates that it is not possible for an organism with a relatively large genome (>100 Mbp) to grow rapidly beyond a few thousand cells in a purely eutelic manner. To grow beyond thousands of cells, development must slow considerably to allow for identification and subsequent repair or destruction of those cells which will inevitably arise with DFS errors. In order to test our theory for larger organisms it is necessary to focus on early stages of development in which rapid coherent DNA replication occurs. The syncytial phase of insects is an important example [2]. Such is the rapidity of replication in this phase, cell division is itself forsaken, with, instead, repeated rounds of nuclear division within the single large cell of the syncytium. Our theory would predict that the number of nuclear divisions is limited according to equation (11).

We test this using data from the most intensively studied insect, the fruitfly *Drosophila melanogaster* [29]. This organism has a haploid genome size of approximately 175 Mbp. In its syncytial phase, it undergoes 13 synchronised rounds of nuclear division, the number of nuclei increasing by a factor of 2 in each round, thus creating approximately 8192 nuclei. Nuclear division then ceases, the nuclei are transported to the syncytial membrane and cellularisation occurs to create the embryonic epiblast. The amount of DNA replicated during the syncytial phase is approximately $8192 \times 2 \times 175\,\text{Mbp} = 2.9\,\text{Tbp}$, which, remarkably, is just below the limit imposed by the universal constant N_U. Interestingly, the haploid mutant (with half as much DNA per nucleus) goes through 14 rounds of nuclear division, resulting in the same amount of DNA being replicated in the syncytium [30]. This could, for example, be explained by the existence of a critical concentration of a key molecule (utilised during replication, and thus being depleted with each round of replication) ensuring that nuclear division in the syncytium does not overstep the N_U bound.

Given that 2.9 Tbp is so close to N_U, a small number of errors will occur with a non-negligible frequency. From equation (8) we see that the probability of having no DFS errors is approximately 38%. A straightforward analysis using Poisson statistics indicates that the probabilities of one and two DFS errors are 37% and 18%, respectively. Thus, fewer than 1 in 10 embryos would have three or more DFS events. The errors can occur in any of the doubling cycles, though will be exponentially more likely to occur in the last few cycles. Presumably, such errors, topologically linking two daughter nuclei, would be left uncorrected with those nuclei excluded from the cellularisation process. One can extend the analysis to catalogue the frequencies with which errors occur in earlier or later cycles, and to then predict the variation in nuclei numbers after 13 cycles, but this lies beyond the scope of the current paper.

One can speculate on the implications of the (diploid) embryo having a hypothetical 14th cycle, thus creating 16 384 nuclei. In this case the amount of DNA to be replicated is almost twice N_U, and fewer than 1 in 6 embryos (15%) would successfully complete the syncytial phase free of DFS errors. Poisson statistics indicate that approximately 1 in 3 embryos (30%) would have three or more errors, and more than 1 in 7 embryos (13%) would accumulate four or more errors. These significantly higher frequencies of error may simply be too costly for robust subsequent development, hence the limitation to 13 cycles of division.

4.3. Highly polyploid cells

We now turn briefly to case II—the significant replication demands in a single cell in which there is a high degree of polyploidy. There are many examples of this phenomenon in the Eukaryota [31], and we examine here three important organismal examples, *Drosophila*, mouse and human, for which high quality data is available.

Many of the cell types in *Drosophila* are polyploid, some highly so [32]. Detailed data are available for three different cell types: fat body cells, midgut cells, and salivary gland cells, and are summarised below in table 2. We note that the product of genome size and ploidy level approaches but does not exceed N_U, indicating that these cells are capable of robust and rapid DNA replication so long as near-saturation of DNA with ROs is utilised. It is striking that mature polyploid cells in *Drosophila* have DNA content limited to a similar degree to the final syncytial phase

Table 2. Data for various cells with high degrees of polyploidy.

Cell type		N_g	M_p	M_pN_g(Tbp)
Drosophila	Fat body	175 Mbp	225	0.039
	Midgut		171	0.030
	Salivary gland		1669	0.29
Rodent TGC		2.7 Gbp	500	1.35
Human megakaryocyte		3.0 Gbp	128	0.38

of the *Drosophila* embryo (both observations consistent with, and possibly linked by, the theory proposed here). A number of studies have reported that the degree of polyploidy is not necessarily constant across the entire genome, with higher rates of ploidy for gene rich regions [33, 34]. As the value of N_t approaches N_U in terminally differentiated endoreplicating cells, one possibility is to tolerate the inevitable DFS errors by allowing deleterious events in regions of the genome which are no longer functionally important. Indeed, under-replicated genomic regions in *Drosophila* polyploid cells suffer from a significant paucity of licensed origins in comparison to those regions rich in active genes [29, 33].

Turning now to mammals, two examples of cell types with very high degrees of polyploidy are trophoblast giant cells (TGCs) (mainly studied in rodents and analogous to cytotrophoblast cells in humans) and megakaryocytes. TGCs are primary cells in placental development [35], while megakaryocytes are the last stage of the differentiation process to produce platelets in the blood [36]. A single megakaryocyte is able to produce several thousand platelets. Both these cell types are large (up to 100 microns in diameter) and use endoreplication to increase their ploidy within a single cell entity. We provide data in table 2, and again we see that these cells have total DNA content that approaches but does not exceed N_U.

5. Discussion

In this paper we have considered extreme safeguards against DFS errors, through a mechanism in which cells saturate their DNA with ROs on a scale of the average nucleosome separation. Using results from our recent theory of DFS statistics we have derived a formula for the probability of DFS error in this case, and find it to be expressed in terms of a fundamental constant $N_U \approx 3 \times 10^{12}$ basepairs, which essentially defines the upper limit of DNA that can be rapidly replicated with minimal chance of DFS error. The constant is fundamental as it arises from a product of two highly conserved biomolecular parameters, cf equations (6) and (7), and is thus expected to be applicable to organisms spanning the Eukaryota.

Our result is particularly relevant to cell division processes which are required to be efficient in time, i.e. in which there is not the leisure of time for costly post-replication repair of DFS errors [8, 37–40]. As such, we have tested our theory against data from developmental processes which require efficient coordinated cell division processes. Our theory suggests there is a hard trade-off between informational complexity (i.e. size of the genome) and architectural complexity (i.e. the number of developmental cell divisions), and that the product of these two be much smaller than N_U. Data from both eutelic organisms and from the *Drosophila* syncytium are in excellent accord with this prediction.

Our theory is also relevant to single cells which have massive DNA content due to high levels of polyploidy. For such cells which are required to replicate their DNA efficiently in time we again expect a trade-off between genome size and degree of polyploidy. Data from three different highly polyploid cell types in *Drosophila*, TGCs in mice, and megakaryocytes in humans are all in accord with the predictions from our theory.

Naturally, none of this constitutes proof that DFS avoidance is the underpinning biological factor in all of these cases. However, the excellent agreement between a diverse range of biological data and our theoretical prediction of the central importance of $N_U \approx 3$ Tbp does strongly suggest the fundamental role of this constant in shaping biological processes in development and polyploidy. Our theory can be tested experimentally by examining cases of developmental failure (or anomalies in polyploid cells) and ascertaining whether these arise from DFS errors.

The hard limit on rapid DNA replication set by N_U suggests, as described by equation (11), that strategies for development must undergo a sharp transition when the product of the number of embryonic cells, M_c and the size of the genome, N_g approaches this value. If the product M_cN_g is well below N_U then the DFS error rate is negligible and there will be no significant bottlenecks to rapid cell division. However, when the product is similar to or greater than N_U, DFS errors are inevitable and the costly repairs thereby required will greatly slow down the developmental process.

One strategy to cope with this limit is simply not to exceed it, and to make every cell and cell division count, i.e. to have a finely choreographed developmental process in which each cell division is pre-programmed. This is eutely, and indeed we find that eutelic organisms across the Eukaryota have very similar genome sizes and cell number counts, respecting the upper bound set by N_U, despite the diverse natural histories and morphologies of the organisms concerned.

The alternative strategy is to divide development into a rapid phase (during which there is a negligible chance of DFS errors) followed, once the product M_cN_g exceeds N_U, by a slower phase (allowing time

for DFS repairs [8, 37–40]). The example of syncytial development in insects appears to be an excellent example: extremely rapid and synchronised nuclear divisions in the syncytium, then slowing to a cellularisation process and subsequent tissue-based gastrulation process. It is remarkable that the *Drosophila* data show that this transition occurs precisely when the N_U limit is reached. The existence of small numbers of polar bodies after syncytial development [41] might indeed correspond to the small number of failed nuclear divisions due to DFS, and Poisson statistics can be employed in conjunction with our theory to provide predictions on the number of polar bodies expected to arise.

Higher organisms have a whole series of developmental transitions related to morphological requirements, e.g. gastrulation, neurulation, limb development [2]. The very first transition from a cluster of cells (i.e. the blastula) to a more structured morphology might be expected to be tuned to N_U, and preliminary data analysis in mammals confirms this. Indeed, equation (8) of our theory provides an estimate of the probability of DFS occurring. If DFS occur during early embryogenesis, and constitute a fatal error, then this estimate provides a lower bound on embryo failure, and work in progress shows these predictions are consistent with data from zebrafish, chicken, and a range of mammalian species [42]. A counterexample is found in the amphibian model *Xenopus*, which undergoes rapid cell division until a cell mass of several thousand cells is formed [2]. As DFS errors will inevitably arise in this case, we postulate that large numbers of cells in the early embryo with DFS errors could be discarded without disruption of future development. This is akin to an *r*-strategy in ecology [43], namely, large numbers of progeny with little parental care and hence high failure rate. In this sense, the choreography of cell division and differentiation in eutelic development is akin to the *K*-strategy, namely, small numbers of progeny with significant parental care to maximise survival.

Returning briefly to the subject of polyploidy, there are extreme cases which break the bound set by N_U. For example, the giant neuronal cells of the sea hare *Aplysia californica* have genome size ∼930 Mbp [44] and ploidy of 600 000 [45] giving a total DNA content of over 500 Tbp. Our theory is moot in such a case, beyond the obvious conclusion that DNA replication in the creation of such cells will be choked with DFS errors requiring repair; other mechanisms beyond replication such as cell fusion may contribute to such enormous ploidy levels. High DFS rates in this case are presumably tolerable for the organism as these cells do not have a role in the earlier developmental processes, and they are not involved in processes which require rapid cell division, unlike the examples we studied earlier, such as trophoblastic giant cells (driving placental development) and megakaryocytes (driving platelet production).

Our theory also places a strict bound on the largest possible genome of an organism, assuming that cell replication must occur with reasonable efficiency at some stages of the organism's life cycle. Assuming a diploid organism, we would predict that one half of N_U, namely 1.5 Tbp, is an upper limit on haploid (or half the value of total) genomic content. This compares favourably with very large genome sizes known in single-celled eukaryotes and plants. Specifically, the estimated genome lengths, found in *Amoeba dubia* and *Amoeba proteus*, are ∼0.67 and ∼0.29 Tbp, respectively [46]. Relatedly, the 2C value for the largest known plant genome, in octaploid *Paris japonica*, corresponds to a genomic content of ∼0.298 Tbp and another very close candidate is the fern *Tmesipteris obliqua* (∼0.294 Tbp) [47].

The question of whether an organism could sustain a 3 Tbp DNA load in each cell is brought into sharp focus by the question: how much volume is required to store the protein complexes needed to saturate such a large genome? Back of the envelope calculations yield some interesting answers. The MCM2-7 double hexamer complex has a volume of approximately 3000 cubic nanometers [48]. To saturate 3 Tbp of DNA (i.e. at the level of one complex per 200 bp repeat of the nucleosome) requires 1.5×10^{10} complexes whose collective volume is therefore approximately 50 000 cubic microns. Interestingly, this is about the size of a large eukaryotic cell (a cell of diameter ∼40 microns). So, it is physically impossible for an organism to achieve saturation of such a large genome without utilising very large cells (particularly in the embryonic stage where errors are presumably less tolerated). It is natural to then ask whether such severe physical constraints are present for the two applications we studied where saturation was assumed, namely eutelic organisms and the *Drosophila* syncytium. The answer is no. The volume of MCM complexes required to saturate the modest 100 Mbp genome of *C. elegans* requires only 0.1% the volume of a typical eukaryotic cell. And the approximately cylindrical *Drosophila* syncytium has length 0.5 mm and diameter 0.15 mm, yielding a volume of $\sim 10^7$ cubic microns. This is ∼200 times larger than the 50 000 cubic microns required to store the MCM complexes necessary to saturate 3 Tbp of DNA.

It may also be interesting to study very large genomes in the Archaea. Here, a modified constant N_U would be required as the details of the molecular machinery for DNA replication and packaging will differ from the Eukaryota. For example, the inter-nucleosome periodicity is ∼140 bp rather than ∼200 bp as in the eukaryotes [49].

We end with a brief comment on the role of fundamental constants in science. The current theoretical framework of physical phenomena involves a small number of fundamental constants. These constants arise from general principles, and are highly valued as conceptual touchstones of physics [50]. Examples are the speed of light in vacuum, c, and Planck's constant, h, which arise, respectively, from relativistic invariance and limits of measurement precision due to quantum uncertainty. From the point of view of biological physics it is tantalising to think that correspondingly general principles exist in living systems, manifesting themselves through fundamental constants. Whether $N_U \approx 3$ Tbp plays such a role, in constraining and guiding developmental strategies of organisms, remains to be seen.

Author contributions

MAM co-designed the project, provided original concepts, performed mathematical calculations, analysed the data, and co-wrote the paper. LA and JJB provided original concepts and co-wrote the paper. TJN co-designed the project, provided original concepts, performed mathematical calculations, analysed the data, and co-wrote the paper.

Acknowledgments

The authors are grateful to Dianbo Liu and Sam Palmer for helpful discussions. MAM, LA and TJN acknowledge prior support from the Scottish Universities Life Sciences Alliance. JJB acknowledges support from Cancer Research UK (Grant C303/A14301) and the Wellcome Trust (Grant WT096598MA). TJN acknowledges prior support from the National Institutes of Health (Physical Sciences in Oncology Centres, U54 CA143682).

Appendix A. Notation used (bp = 'basepair', and bp^{-1} = 'per basepair')

- N_U fundamental constant (≈ 3 Tbp)
- U inverse of N_U ($\approx 3.3 \times 10^{-13}$ bp^{-1})
- N_t total length of DNA to be replicated in a given biological process
- N_g length of DNA in the haploid genome of a given organism
- N_n length of DNA in one period of the nucleosome repeat (≈ 200 bp)
- N_l average separation between ROs
- N_s median stalling distance of the replication machinery ($\approx$12 Mbp)
- q per nucleotide probability of fork stall ($= \ln(2)/N_\text{s} \approx 5.8 \times 10^{-8}$ bp^{-1})
- ρ probability of a given inter-nucleosome region being occupied by an RO
- R coefficient of variation of RO positions
- α the numerical constant $(\ln 2)^2/2 \approx 0.240$
- M_p degree of polyploidy (equal to unity for haploid cells)
- M_c number of cells after a given sequence of coordinated cell divisions

Appendix B. Derivation of the central result for arbitrary values of ρ

In the recent theory of DFS [6], an expression was derived for the probability of one or more DFS during the replication of a genome. The expression is valid for an arbitrary distribution of ROs so long as all inter-RO distances are much smaller than the median stalling distance, N_s. Translating that result into the notation used in this paper, we have:

$$P_\text{error}(N_\text{t}) = 1 - \exp\left(-\alpha \frac{N_l N_\text{t}}{N_\text{s}^2}\left(1 + R^2\right)\right). \quad \text{(Ai)}$$

The task is to implement this formula for the quantised situation in which there is a density ρ of ROs occupying inter-nucleosome regions of periodicity N_n. This requires us to calculate the mean spacing of ROs, N_l, and the coefficient of variation R.

Both of these can be obtained in a straightforward manner as follows. Subsequent calculations use the nucleosome periodicity N_n as a unit of DNA length. The probability of a gap of k units is $\rho(1-\rho)^{k-1}$. Thus, the mean RO separation is N_n multiplied by the first moment of this distribution:

$$N_l = N_\text{n}\langle k\rangle = N_\text{n}\sum_{k=1}^{\infty} k\rho(1-\rho)^{k-1} = N_\text{n} \,/\, \rho. \quad \text{(Aii)}$$

The second moment of the distribution is given by

$$\left\langle k^2\right\rangle = \sum_{k=1}^{\infty} k^2\rho(1-\rho)^{k-1} = (2-\rho) \,/\, \rho^2. \quad \text{(Aiii)}$$

Thus, the square of the coefficient of variation is given by

$$R^2 = \frac{\left\langle k^2\right\rangle - \langle k\rangle^2}{\langle k\rangle^2} = 1 - \rho. \quad \text{(Aiv)}$$

Substituting these expressions into (Ai), we have

$$P_\text{error}(N_\text{t}) = 1 - \exp\left(-\alpha \frac{N_\text{n} N_\text{t}}{N_s^2}\frac{(2-\rho)}{\rho}\right). \quad \text{(Av)}$$

In terms of the fundamental constant, the central result for arbitrary ρ is:

$$P_\text{error}(N_\text{t}) = 1 - \exp\left(-\frac{N_\text{t}}{N_U}\frac{(2-\rho)}{\rho}\right), \quad \text{(Avi)}$$

which is given in the main text as equation (10) and which reduces to equation (8) for the case $\rho \to 1$.

Appendix C. Further analysis of the central result for $\rho < 1$ and the derivation of a bound on ρ

Denoting the $\rho = 1$ (saturated) form of equation (Avi) by $P_{\text{error}}^{\text{sat}}$ we have

$$P_{\text{error}}^{\text{sat}} = 1 - \exp\left(-\frac{N_{\text{t}}}{N_U}\right), \qquad \text{(Avii)}$$

and then equations (Avi) and (Avii) can be combined to give the relationship

$$P_{\text{error}}\left(N_{\text{t}}\right) = 1 - \left(1 - P_{\text{error}}^{\text{sat}}\right)^{\frac{(2-\rho)}{\rho}}. \qquad \text{(Aviii)}$$

This useful relationship allows a direct calculation of the probability of DFS errors with non-saturated coverage of ROs (i.e. $\rho < 1$) from the probability of DFS errors with saturated coverage. For $P_{\text{error}}^{\text{sat}} \ll 1$ the relationship simplifies dramatically to

$$P_{\text{error}}\left(N_{\text{t}}\right) \approx \frac{(2-\rho)}{\rho} P_{\text{error}}^{\text{sat}}. \qquad \text{(Aix)}$$

This shows a rapid increase in the DFS error rate as coverage reduces from $\rho = 1$. For example, the increase in error rate is threefold when $\rho = 1/2$ and roughly 20-fold when $\rho = 1/10$.

A lower bound on ρ can be derived from equation (Aviii) which might prove useful when comparing our theory with more detailed experimental data. The failure rate of embryogenesis will be due to a range of factors, including, we argue severe DNA replication errors such as DFS. Thus, if we denote by P_{obs} the experimentally observed embryo failure rate, we have $P_{\text{obs}} \geqslant P_{\text{error}}$. Combining this inequality with equation (Aviii) and taking logarithms to isolate ρ gives, after some manipulation, the inequality:

$$\rho \geqslant \frac{\log\left(\left(1 - P_{\text{error}}^{\text{sat}}\right)^2\right)}{\log\left(\left(1 - P_{\text{error}}^{\text{sat}}\right)\left(1 - P_{\text{obs}}\right)\right)}. \qquad \text{(Ax)}$$

If all the error rates are much smaller than unity this expression simplifies greatly to

$$\rho \geqslant \frac{2P_{\text{error}}^{\text{sat}}}{\left(P_{\text{error}}^{\text{sat}} + P_{\text{obs}}\right)}. \qquad \text{(Axi)}$$

These expressions are discussed further in the main text following equation (12).

ORCID iDs

L Albergante https://orcid.org/0000-0001-8151-6989
J J Blow https://orcid.org/0000-0002-9524-5849
T J Newman https://orcid.org/0000-0002-0332-337X

References

[1] Alberts B, Johnson A, Lewis J, Raff M, Roberts K and Walter P 2002 *Molecular Biology of the Cell* (New York: Garland Science)
[2] Wolpert L, Beddington R, Jessell T, Lawrence P, Meyerowitz E and Smith J 2002 *Principles of Development* (Oxford: Oxford University Press)
[3] McIntosh D and Blow J J 2012 Dormant origins, the licensing checkpoint, and the response to replicative stresses *Cold Spring Harbor Perspect. Biol.* **4** a012955
[4] Blow J J and Ge X Q 2009 A model for DNA replication showing how dormant origins safeguard against replication fork failure *EMBO Rep.* **10** 406–12
[5] Blow J J and Dutta A 2005 Preventing re-replication of chromosomal DNA *Nat. Rev. Mol. Cell Biol.* **6** 476–86
[6] Newman T J, Mamun M A, Nieduszynski C A and Blow J J 2013 Replisome stall events have shaped the distribution of replication origins in the genomes of yeasts *Nucleic Acids Res.* **41** 9705–18
[7] Al Mamun M, Albergante L, Moreno A, Carrington J T, Blow J J and Newman T J 2016 Inevitability and containment of replication errors for eukaryotic genome lengths spanning megabase to gigabase *Proc. Natl Acad. Sci. USA* **113** E5765–74
[8] Moreno A, Carrington J T, Albergante L, Al Mamun M, Haagensen E J, Komseli E-S, Gorgoulis V G, Newman T J and Blow J J 2016 Unreplicated DNA remaining from unperturbed S phases passes through mitosis for resolution in daughter cells *Proc. Natl Acad. Sci. USA* **113** E5757–64
[9] Karschau J, Blow J J and de Moura A P S 2012 Optimal placement of origins for DNA replication *Phys. Rev. Lett.* **108** 058101
[10] Gauthier M G, Norio P and Bechhoefer J 2012 Modeling inhomogeneous DNA replication kinetics *PLoS One* **7** e32053
[11] Rhind N and Gilbert D M 2013 DNA replication timing *Cold Spring Harbor Perspect. Biol.* **5** a010132
[12] Blow J J, Gillespie P J, Francis D and Jackson D A 2001 Replication origins in *Xenopus* egg extract are 5–15 kilobases apart and are activated in clusters that fire at different times *J. Cell Biol.* **152** 15–26
[13] Mahbubani H M, Chong J P J, Chevalier S, Thömmes P and Blow J J 1997 Cell cycle regulation of the replication licensing system: involvement of a Cdk-dependent inhibitor *J. Cell Biol.* **136** 125–35
[14] Smits A H, Lindeboom R G H, Perino M, van Heeringen S J, Veenstra G J C and Vermeulen M 2014 Global absolute quantification reveals tight regulation of protein expression in single *Xenopus* eggs *Nucleic Acids Res.* **42** 9880–91
[15] Wühr M, Freeman R M, Presler M, Horb M E, Peshkin L, Gygi S and Kirschner M W 2014 Deep proteomics of the *Xenopus laevis* egg using an mRNA-derived reference database *Curr. Biol.* **24** 1467–75
[16] Woodward A M, Göhler T, Luciani M G, Oehlmann M, Ge X, Gartner A, Jackson D A and Blow J J 2006 Excess Mcm2-7 license dormant origins of replication that can be used under conditions of replicative stress *J. Cell Biol.* **173** 673–83
[17] Blow J J and Gillespie P J 2020 Density of dormant origins in the *Xenopus* embryo (in preparation)
[18] Bradbury E M and Van Holde K E 1989 *Chromatin Series in Molecular Biology* (New York: Springer)
[19] Szerlong H J and Hansen J C 2011 Nucleosome distribution and linker DNA: connecting nuclear function to dynamic chromatin structure *Biochem. Cell Biol.* **89** 24–34
[20] Remus D, Beuron F, Tolun G, Griffith J D, Morris E P and Diffley J F X 2009 Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing *Cell* **139** 719–30

[21] Evrin C, Clarke P, Zech J, Lurz R, Sun J, Uhle S, Li H, Stillman B and Speck C 2009 A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication *Proc. Natl Acad. Sci. USA* **106** 20240–5
[22] van Cleave H J 1932 Eutely or cell constancy in its relation to body size *Q. Rev. Biol.* **7** 59–67
[23] Coghlan A 2005 Nematode genome evolution *WormBook* ed The *C. elegans* Research Community (https://www.wormbook.org)
[24] Gabriel W N, McNuff R, Patel S K, Gregory T R, Jeck W R, Jones C D and Goldstein B 2007 The tardigrade *Hypsibius dujardini*, a new model for studying the evolution of development *Dev. Biol.* **312** 545–59
[25] Yoshida Y *et al* 2017 Comparative genomics of the tardigrades *Hypsibius dujardini* and *Ramazzottius varieornatus PLoS Biol.* **15** e2002266
[26] Wallace R L 2002 Rotifers: exquisite metazoans *Integr. Comp. Biol.* **42** 660–7
[27] Kim H-S *et al* 2018 The genome of the freshwater monogonont rotifer *Brachionus calyciflorus Mol. Ecol. Resour.* **18** 646–55
[28] Rodríguez-Martínez M, Pinzón N, Ghommidh C, Beyne E, Seitz H, Cayrou C and Méchali M 2017 The gastrula transition reorganizes replication-origin selection in *Caenorhabditis elegans Nat. Struct. Mol. Biol.* **24** 290–9
[29] Hua B L and Orr-Weaver T L 2017 DNA replication control during *Drosophila* development: insights into the onset of S phase, replication initiation, and fork progression *Genetics* **207** 29–47
[30] Edgar B A, Kiehle C P and Schubiger G 1986 Cell cycle control by the nucleo-cytoplasmic ratio in early Drosophila development *Cell* **44** 365–72
[31] Orr-Weaver T L 2015 When bigger is better: the role of polyploidy in organogenesis *Trends Genet.* **31** 307–15
[32] Nordman J, Li S, Eng T, Macalpine D and Orr-Weaver T L 2011 Developmental control of the DNA replication and transcription programs *Genome Res.* **21** 175–81
[33] Sher N, Bell G W, Li S, Nordman J, Eng T, Eaton M L, Macalpine D M and Orr-Weaver T L 2012 Developmental control of gene copy number by repression of replication initiation and fork progression *Genome Res.* **22** 64–75
[34] Frawley L E and Orr-Weaver T L 2015 Polyploidy *Curr. Biol.* **25** R353–8
[35] MacAuley A, Cross J C and Werb Z 1998 Reprogramming the cell cycle for endoreduplication in rodent trophoblast cells *Mol. Biol. Cell* **9** 795–807
[36] Machlus K R and Italiano J E 2013 The incredible journey: From megakaryocyte development to platelet formation *J. Cell Biol.* **201** 785–96
[37] Minocherhomji S, Ying S, Bjerregaard V A, Bursomanno S, Aleliunaite A, Wu W, Mankouri H W, Shen H, Liu Y and Hickson I D 2015 Replication stress activates DNA repair synthesis in mitosis *Nature* **528** 286–90
[38] Bhowmick R, Minocherhomji S and Hickson I D 2016 RAD52 facilitates mitotic DNA synthesis following replication stress *Mol. Cell* **64** 1117–26
[39] Fragkos M and Naim V 2017 Rescue from replication stress during mitosis *Cell Cycle* **16** 613–33
[40] Spies J, Lukas C, Somyajit K, Rask M-B, Lukas J and Neelsen K J 2019 53BP1 nuclear bodies enforce replication timing at under-replicated DNA to limit heritable DNA damage *Nat. Cell Biol.* **21** 487–97
[41] Williams B C, Dernburg A F, Puro J, Nokkala S and Goldberg M L 1997 The Drosophila kinesin-like protein KLP3A is required for proper behavior of male and female pronuclei at fertilization *Development* **124** 2365–76
[42] Al Mamun M 2019 Fundamental limit on genomic information regulates developmental success in eukaryotes (in preparation)
[43] MacArthur R H and Wilson E O 1967 *The Theory of Island Biogeography* (Princeton, NJ: Princeton University Press)
[44] Lasek R J and Dower W J 1971 Aplysia californica: analysis of nuclear DNA in individual nuclei of giant neurons *Science* **172** 278–80
[45] Moroz L L 2011 *Aplysia Curr. Biol.* **21** R60–1
[46] Parfrey L W, Lahr D J and Katz L A 2008 The dynamic nature of eukaryotic genomes *Mol. Biol. Evol.* **25** 787–94
[47] Hidalgo O, Pellicer J, Christenhusz M, Schneider H, Leitch A R and Leitch I J 2017 Is there an upper limit to genome size? *Trends Plant Sci.* **22** 567–73
[48] Li N, Zhai Y, Zhang Y, Li W, Yang M, Lei J and Gao N 2015 Structure of the eukaryotic MCM complex at 3.8 Å *Nature* **524** 186–91
[49] Mattiroli F *et al* 2017 Structure of histone-based chromatin in Archaea *Science* **357** 609–12
[50] Leggett A J 1987 *The Problems of Physics* (Oxford: Oxford University Press)