Mathematical model for the distribution of DNA replication origins

Alessandro de Moura[1,*] and Jens Karschau[1]

[1] *Institute for Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3UE, United Kingdom*

(Dated: May 21, 2024)

DNA replication in yeast and in many other organisms starts from well-defined locations on the DNA known as replication origins. The spatial distribution of these origins in the genome is particularly important in ensuring that replication is completed quickly. Cells are more vulnerable to DNA damage and other forms of stress while they are replicating their genome. This raises the possibility that the spatial distribution of origins is under selection pressure. In this work we investigate the hypothesis that natural selection favours origin distributions leading to shorter replication times. Using a simple mathematical model, we show that this hypothesis leads to two main predictions about the origin distributions: that neighbouring origins that are inefficient (less likely to fire) are more likely to be close to each other than efficient origins; and that neighbouring origins with larger differences in firing times are more likely to be close to each other than origins with similar firing times. We test these predictions using next-generation sequencing data, and show that they are both supported by the data.

I. INTRODUCTION

The life cycle of an Eukaryotic unicellular organism culminates on the S-phase, when DNA replication takes place. The cell can only proceed to cell division once all DNA in the cell has been fully replicated. This is a crucial point in the cell's life, and there are strong selection pressures to ensure that DNA replication is as fast and error-free as possible [1, 2].

Replication starts from locations in the genome called *replication origins*[3–7]. In many organisms, the locations of replication origins are fixed, and are determined by specific DNA motifs called ARS consensus sequences (ARS stands for Autonomous Replicating Origin). This is the case with most unicellular Eukaryotes, including *S. cerevisiae*, the model organism we focus on in this paper.

A typical Eukaryotic organism has many replication origins. *S. cerevisiae*, for example, has 459 origins distributed throughout 16 chromosomes. This means that chromosomes have multiple origins, with larger chromosomes having more origins.

Before it can start replication, an origin must have been *licensed* before S-phase starts [7–12]. In *S. cerevisiae*, origin licensing is initiated by the binding of the origin recognition complex (ORC) to the origin site. In the next step, the proteins Cdc6 and Cdt1 bind to ORC, forming a complex which then loads copies of the Mcm2-Mcm7 hexamer ring (MCM) and clamps them around the DNA molecule at the origin site [13], thereby completing the licensing process. Only licensed origins will be able to initiate replication during S-phase.

Once S-phase starts, licensing stops — depending on the organism, the licensing proteins are detached from DNA and then either degraded or inactivated [14, 15]. But some origins in any given cell will fail to license before the S-phase. Success or failure of licensing any given origin is a stochastic process: in a population of genetically identical cells, one cell may fail to license some origin, while another cell in the population may succeed in licensing that same origin [16]. This naturally leads to the definition of an origin's *competence*, which is the probability p_i of origin i successfully licensing [16]. Alternatively, p_i is the fraction of cells in a population where origin i is licensed in a replication round.

Measuring origin competences for a whole genome is a difficult task. The duplication rate of plasmids incorporating an ARS in principle allows one to measure the competence of chosen origins [17, 18]. It would be extremely laborious, however, to use this method to obtain competences for a whole genome, and the results are not always reliable. Next-generation sequencing makes it possible to measure replication profiles with unprecedented resolution, enabling one to take genome-wide snapshots at controlled times of the state of replication of a population of cells [19]. Fitting a detailed computational model of DNA replication in yeast to this data resulted in a reliable estimation of fundamental origin parameters such as competence and mean firing time for every origin in yeast [16, 20–23]. This allows us to ask questions about how the positions of origins may be related to their firing dynamics.

The cell becomes extremely vulnerable once DNA replication starts: DNA damage caused by UV radiation and by other stresses is much more likely to kill the cell in this stage of its life cycle [24, 25], due to the induction of fork stalling and collapse. So it is in the cell's best interests to shorten the time it spends replicating its DNA. And we know that replication origins are often created and destroyed throughout evolution [26]. This motivated us to propose, in a previous work [27], that natural selection favours origin distributions resulting in shorter replication times. We showed that this hypothesis leads to the prediction that there is a correlation between the positions of origins and their competences: neighbouring origins with low competences are expected to be located

* a.moura@abdn.ac.uk

close to each other, while origins with high competence are expected to be far away from other origins. However, no genome-wide estimation of origin competences was available when that work was published; so no experimental validation was possible, and that was a mostly theoretical work.

In this paper, we use the origin parameters derived from next-generation sequencing data to quantitatively test the hypothesis proposed in [27], showing that it is supported by the data. We also extend the model presented in [27] in a major way, by incorporating the mean firing times of origins into the model. This enables us to predict a correlation between the firing times of neighbouring origins and their genomic distance. This new prediction is tested using two different next-generation sequencing data sets, and we show that it also agrees with the data, adding evidence to the idea that origin locations have been selected by evolution to favour short replication times.

II. THE MODEL

The goal of the mathematical model we formulate here is to understand how the spatial distribution of replication origins affects the replication time; in particular, we are interested in the origin configurations that lead to the shortest replication times. We consider the simplest possible case of a chromosome with only two origins. Even though this is a very idealised model, we argue that it captures some essential aspects of how the replication time depends on the positioning of origins. The goal is to apply insights gained from this model to analyse real replication data from yeast. We generalise here the analysis done in [27], extending that model to take into account the different firing times of origins.

S. cerevisiae and Eukaryotes in general have linear chromosomes — as opposed to the circular chromosomes found in bacteria. So the model we develop in the following assumes linear chromosomes.

In order to simplify the model, we choose a unit of length such that the chromosome length is 1. The positions of the two origins are denoted by x_1 and x_2, with $0 \leq x_1 \leq x_2 \leq 1$. The competences of the two origins are denoted by p_1 and p_2. The last parameter characterising the model is the difference in the firing time between the two origins, denoted by τ. We assume that origin 1 is the early origin — that is, origin 1 never fires later than origin 2. This is done for mathematical convenience, and it does not affect the generality of our conclusions in any way. The replication forks are assumed to travel with constant velocity v, which is consistent with recent fork velocity profiles [28].

We choose $t = 0$ as the moment when origin 2 fires. Assuming both origins fire, the order of events is therefore as follows: at time $t = -\tau$, origin one fires; at time $t = 0$, origin 2 fires; and then, the whole chromosome is replicated after all forks have either collided with other forks or reached one of the ends of the chromosome. Notice that if a fork reaches an origin before it replicates, the fork goes through the origin and continues in its way; the origin is said to have been "passively replicated" by the fork.

This process is illustrated in Fig 1(a). We assume that $\tau \geq 0$; that is, origin 1 either fires earlier than origin 2, or fires at the same time as origin 2. τ is therefore the difference in firing times of the two origins. For the purposes of this simple model, we assume that the firing times of both origins are constant, and subject to no stochastic variation. In reality, origin firing times are stochastic [16, 21, 22, 29, 30], but we ignore this at first for the sake of simplicity. We do investigate the case of stochastic firing later on in this paper.

A. Configurations of minimum replication time

We now proceed to determine where the origins should be placed so that the replication time is shortest. Let us first assume that both origins successfully fire (that is, both have been licensed). Let us focus on the moment when origin 2 (the later origin) fires, at a time τ after origin 1 has fired. At this time, a portion of the chromosome has already been replicated by the forks created at origin 1, as indicated by the thicker lines in Fig. 1. The question is then how to place the origins so that the remaining unreplicated portion of the chromosome is replicated as fast as possible.

The first step is to notice that after origin 2 fires, there are four forks travelling on the chromosome, and so the rate with which replication is proceeding is $4v$ (in replicated length per second). This will remain the replication rate as long as there are four forks on the chromosome. If, for example, origin 2 is very close to the right edge of the chromosome, its right-propagating fork will hit the edge and disappear soon after origin 2 fires. In that case, only three forks will remain, and replication will take place at a lower rate than the maximum possible rate, because of the absence of one fork.

The conclusion is that in order to get the shortest possible replication time, all four forks must coexist until the end of replication, so that replication remains at its fastest rate until it finishes. In other words, in the optimal configuration all forks must terminate simultaneously. This implies that the origin positions must be such that each fork will travel one-fourth of the remaining chromosomal length yet to be replicated. This reasoning leads to the origin placement shown in Fig. 1(b), where δ is defined as one-quarter of the unreplicated length of chromosome at the time origin 2 fires. Notice that in this configuration, the unreplicated region between the two origins is twice as large as the regions next to the edges. The reason is that the central region will be replicated by two forks, whereas the left and right regions are replicated by only one fork each.

The above argument assumed both origins successfully

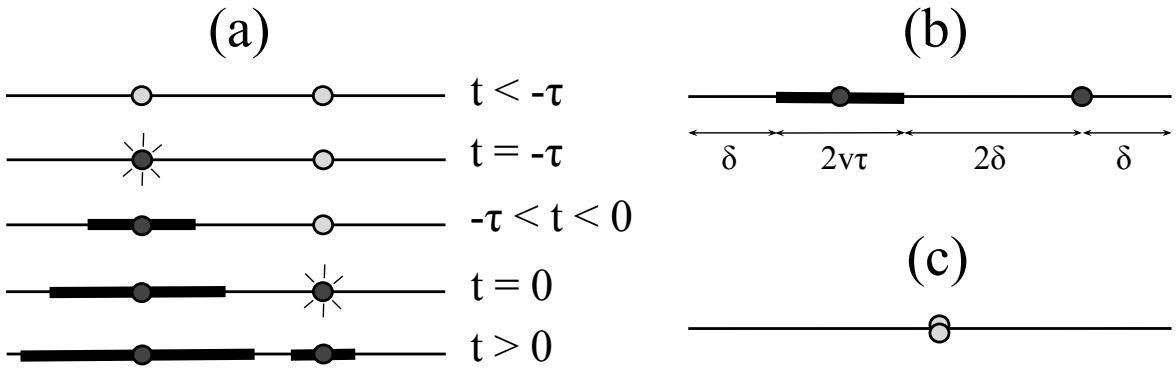


FIG. 1. Depiction of idealised two-origin chromosome. **(a)** Sequence of origin firings in the two-origin chromosome of the model. The horizontal lines represent the chromosome, and the thick lines are the regions that have been replicated. Time progresses from top to bottom. **(b)** Minimum replication time chromosome, assuming both origins fire, depicted at $t = 0$. **(c)** Clustered origin configuration, with both origins located in the centre.

fired. In reality, however, each origin has a probability of firing, given by its competence. The natural quantity characterising replication time is therefore the mean replication time, which we denote by $T_{\rm av}^{\rm sep}$. It takes into account the probabilities of all the possible combinations of origins firing and failing (we will discuss later how we deal with the awkward case of none of the origins firing.)

If both origins always fire, Fig. 1(b) is the configuration with the least average replication time. If the origins have a substantial probability of failing, however, the configuration shown in Fig. 1(b) will no longer be the one with the least replication time. The reason is simple: if one of the origins fails to fire, the origin that does fire is closer to one end of the chromosome than the other, and so one fork will terminate before the other. This violates the condition for minimal replication time, as explained above. If we knew ahead of time that only one origin would fire, the origin placement leading to the shortest replication time would be the clustered configuration depicted in Fig. 1(c), with both origins located at the centre of the chromosome. This configuration ensures that both forks created by the one firing origin terminate simultaneously. We thus expect to see a transition in the optimal origin configuration (that is, the one with the least replication time) from Fig. 1(b) for high competences to Fig. 1(c) for low competences.

Consider now the effect that the difference in firing times has on the optimal configuration, assuming for the purposes of this discussion that the origins have high competences. If the two origins fire at almost the same time, then from the arguments presented above the configuration in Fig. 1(b) is the one with minimal replication time. If, however, one of the origins is much earlier than the other, so that the later origin is passively replicated by a fork from the earlier origin before it has a chance of firing, this changes. In that case, replication will mostly be done by the earlier origin, and for the purposes of computing the replication time, it is as if the later origin did not exist. By this argument, the case of a large firing time difference τ is equivalent to the case of one single firing origin in the previous paragraph, and we expect the clustered configuration of Fig. 1(c) to be the one with the shortest replication time. We therefore expect to see a transition in the shortest-time configuration, from isolated to clustered origins, as τ increases.

We now proceed to investigate these transitions in detail through a simple mathematical model.

B. The case of separated origins

First, we assume that both origins have competence 1, and never fail to fire. As explained above, in this case the configuration of minimum replication time is described by Fig. 1b. By time $t = 0$, each fork that started at origin 1 (the earlier-firing origin) will have travelled for time τ with velocity v. Therefore, the length of the replicated region at time $t = 0$ is $2v\tau$; see Fig. 1(b). So at $t = 0$, the total length of the unreplicated part of the chromosome is $1-2v\tau$. Therefore, the quantity δ in Fig. 1(b) is given by

$$\delta = \frac{1-2v\tau}{4} = \frac{1}{4} - \frac{v\tau}{2}. \tag{1}$$

The locations of the two origins in Fig. 1(b) are thus

$$x_1 = \delta + v\tau = \frac{1}{4} + \frac{v\tau}{2}; \qquad x_2 = 1 - \delta = \frac{3}{4} + \frac{v\tau}{2}. \tag{2}$$

If the origins fire simultaneously, the time difference τ is zero. In this particular case, we find from Eq. (2) that the locations of the origins which minimise replication time are $x_1 = 1/4$ and $x_2 = 3/4$, where they are positioned symmetrically around the centre of the chromosome. We thus recover the result presented in [27], which assumes simultaneous firing of origins. The novel feature of our more general model is in the terms containing τ in Eq.

(2). They show that the effect of non-simultaneous firing is to break this symmetry and displace both origins so that the earlier origin (origin 1) moves closer to the centre of the chromosome, while preserving its distance from the later origin (origin 2).

The chromosome shown in Fig. 1(b) will finish replicating when each fork traverses a distance of δ. Since our "clock" starts when origin 2 fires, this replication time is

$$T_{12}^{\rm sep} = \delta/v = \frac{1}{4v} - \frac{\tau}{2}. \tag{3}$$

The "12" subscript indicates that this is the replication time assuming both origins fire, and "sep" indicates that we are considering the case where the origins are spatially separated.

Now let us drop the assumption that origins have competence 1. If the origin locations are still as shown in Fig. 1(b), but now origin 2 fails to fire, the new replication time is

$$T_{1}^{\rm sep} = 3\delta/v = \frac{3}{4v} - \frac{3\tau}{2}, \tag{4}$$

since the entire region to the right of origin 1 is replicated by the right-propagating fork originated at origin 1, which must traverse a distance of 3δ to reach the end of the chromosome. Similarly, if origin 1 fails, the region to the left of origin 2 is now entirely replicated by the fork originating there, and the corresponding replication time is

$$T_{2}^{\rm sep} = \frac{3\delta + 2v\tau}{v} = T_{1}^{\rm sep} + 2\tau. \tag{5}$$

Using the definitions of the origins' competences, p_1 and p_2, we can write the expression for the average replication time $T_{\rm av}^{\rm sep}$:

$$\begin{aligned} T_{\rm av}^{\rm sep} = \; & p_1 p_2 T_{12}^{\rm sep} + p_1(1-p_2)T_{1}^{\rm sep} + (1-p_1)p_2 T_{2}^{\rm sep} \\ & + (1-p_1)(1-p_2)T_{\rm fail}. \end{aligned} \tag{6}$$

The first three terms on the right-hand side are the contributions to $T_{\rm av}^{\rm sep}$ from the three scenarios described above: both origins firing; only origin 1 firing; and only origin 2 firing. The fourth term takes care of the case when both origins fail. That case would correspond to an undefined (or infinite) replication time, and it would make the definition of $T_{\rm av}^{\rm sep}$ awkward. So in order to make $T_{\rm av}^{\rm sep}$ well-defined, we postulate that the case where both origins fail corresponds to some constant replication time, which we denote by $T_{\rm fail}$. This is justifiable because this is not meant to be a fully realistic model of DNA replication in yeast; it is rather meant to capture fundamental aspects of the replication dynamics. In real organisms with many origins per chromosome, the region near two unlicensed origins will eventually be replicated by forks originated from other origins. Hence, it makes sense to use a finite value for $T_{\rm fail}$. The precise value is not needed for any of the predictions we make from the model: we shall see shortly that $T_{\rm fail}$ cancels out in all the relevant equations.

Substituting Eqs. (3), (4) and (5) into Eq. (6), we get

$$\begin{aligned} T_{\rm av}^{\rm sep} = \; & \tfrac{1}{4v}\left(2p_1p_2v\tau - 5p_1p_2 - 6p_1v\tau + 3p_1 + 2p_2v\tau + 3p_2\right) \\ & + (1-p_1)(1-p_2)T_{\rm fail} \end{aligned} \tag{7}$$

C. The case of clustered origins.

As argued above, it is reasonable to expect that if the origins have low enough competences, the configuration with both origins clustered at the centre of the chromosome (see Fig. 1c) will have the shortest replication time. We now proceed to find expressions for the replication time of the clustered case, mirroring what we did above.

Since both origins now occupy the same position, if the earlier origin (i.e. origin 1) fires, that corresponds to both $T_{1}^{\rm clu}$ and $T_{12}^{\rm clu}$ — since origin 2 is then passively replicated by the forks created by the earlier origin. Both forks traverse half the length of the chromosome, and therefore they take the time $1/2v$ to finish. But they started at time $t = -\tau$, so we have

$$T_{1}^{\rm clu} = T_{12}^{\rm clu} = \frac{1}{2v} - \tau. \tag{8}$$

If the earlier origin fails and the later origin fires, we have instead

$$T_{2}^{\rm clu} = \frac{1}{2v}. \tag{9}$$

This leads to the expression for the average replication time for the clustered configuration:

$$T_{\rm av}^{\rm clu} = \frac{1}{2v}\left(p_1 + p_2 - p_1p_2 - 2p_1v\tau\right) + (1-p_1)(1-p_2)T_{\rm fail}. \tag{10}$$

D. Transitions in the shortest-time origin configuration.

As the competences of origins decrease, we expect to see a transition of the configuration with the shortest replication time, from one with separate origins to one where the origins are clustered. That is, for high p_1 and p_2, we expect $T_{\rm av}^{\rm sep} < T_{\rm av}^{\rm clu}$, whereas for low p_1 and p_2, we expect $T_{\rm av}^{\rm clu} < T_{\rm av}^{\rm sep}$. The transition point is therefore determined by $T_{\rm av}^{\rm sep} - T_{\rm av}^{\rm clu} = 0$. Substituting Eqs (7) and (10), and solving for p_2, we find

$$p_2 = \frac{p_1\left(2v\tau - 1\right)}{2p_1v\tau - 3p_1 + 2v\tau + 1}. \tag{11}$$

Notice that the terms involving $T_{\rm fail}$ drop out; as promised, $T_{\rm fail}$ has no effect on the conclusions that follow.

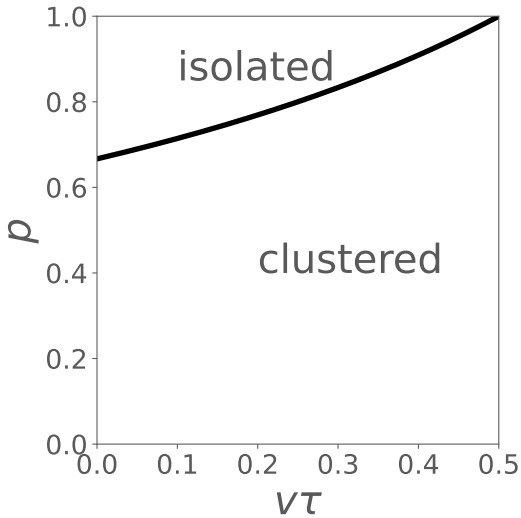


FIG. 2. Phase diagram for the optimal time configuration of two-origin chromosomes. The horizontal axis is the difference in replication times between their origins, and the vertical axis is their competences. Above the black line, isolated origins yield minimal replication time; below it, a clustered configuration with origins very close together is optimal.

An interesting case is when $p_1 = p_2 \equiv p_c$. Then Eq. (11) becomes an equation for p_c and τ. Solving for p_c, we get

$$p_c = \frac{2}{3 - v\tau}. \tag{12}$$

For $p < p_c$, the clustered configuration is the one with the shortest replication time, whereas for $p > p_c$ separate origins replicate the chromosome faster. Eq. (12) shows that p_c increases as τ increases, and hence greater differences in the firing times of neighbouring origins favour the clustering of origins. Eq. (12) can be visualised as the phase diagram shown in Fig. 2, depicting the division of the (τ, p) parameter space into a region where clustered origins or isolated origins yield minimal replication times. For the particular case $\tau = 0$, we recover the result $p_c = 2/3$ found in [27].

An important point is that the transition between the clustered and separated states is always abrupt — the origins do not gradually approach each other as p decreases. Instead, in the minimum replication time configuration, they are either in the configuration shown in Fig. 1(b) or in the clustered state shown in Fig. 1(c); there is no intermediate state. The reason for this is explained in [27]. We confirm this through the numerical simulations described in the following.

E. Numerical simulations

In order to confirm the transitions between the clustered and isolated configurations predicted above, we ran extensive numerical simulations of the replication dynamics.

For a given origin configuration (x_1, x_2), the program generates a population of "virtual chromosomes". For each member of the population, a random number r between 0 and 1 is generated for each origin, and then compared to that origin's competence p_i; if $r < p_i$, the origin is considered to have successfully licensed. Then based on which origins fired, and the firing times and positions of the origins, the replication time is computed. By averaging the firing times over the virtual population, the average replication time for the given origin positions, competences and firing times are found. For given competences p_1 and p_2, the optimal locations of the origins are found by computing the average replication time on a 200×200 grid on the (x_1, x_2) space, and choosing the point on the grid yielding the minimum replication time. For convenience, we use the constraint $x_1 \leq x_2$; this takes advantage of the symmetry of the system to halve the number of points on the grid we need to consider. The size of the virtual population was 10000. The fork velocity is fixed throughout at $v = 1$, and for simplicity the origins were assumed to have the same potential $p = p_1 = p_2$.

We ran simulations for both constant and stochastic firing times. In the simulations using stochastic firing times, after determining whether each origin has been licensed, its firing time is chosen from a normal distribution. The mean of the normal distribution plays the role of the origin's replication time in this version of the model. The standard deviation of the firing time distributions is constant and equal to 0.1.

The results for constant firing times are shown in Figs. 3a and 3b; the results with stochastic firing times are displayed in Figs. 3c and 3d. It is clear that transitions from clustered to isolated configurations take place as p increases (Figs. 3a and 3c), and as τ decreases (Figs. 3b and 3d), as predicted by our theory. For constant firing times, we confirmed that the values of the parameters where the transition takes place match the values predicted by the theory. And although we do not predict analytically the transition point for the stochastic case, the transition still happens, in the direction predicted by the model. This shows that the clustered-isolated transition is robust, and is present regardless of the details of the model. This gives us some confidence that our model captures fundamental aspects of the replication dynamics.

F. Predictions inferred from the model

When using the results of the analysis of the replication dynamics of the idealised two-origin chromosome to

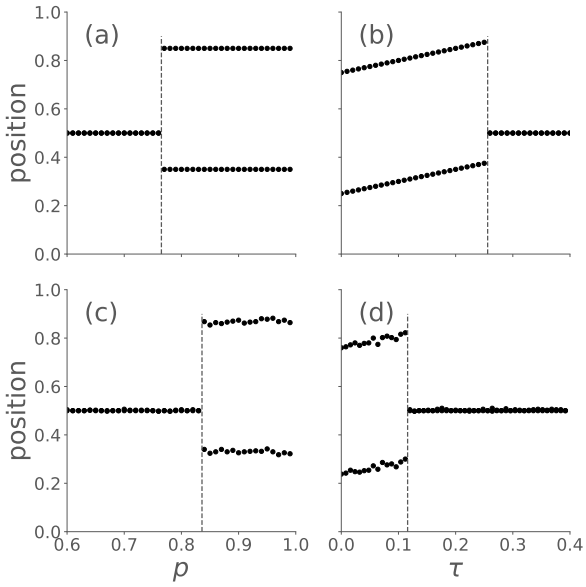


FIG. 3. Simulations of the optimal configurations of origin positions in an idealised two-origin chromosome. $v = 1$ throughout, and the two origins have the same potential, $p = p_1 = p_2$. **(a)** Optimal positions versus the potential p, for a fixed firing time difference $\tau = 0.2$, and constant firing times. **(b)** Optimal positions versus the firing time difference τ, for a fixed potential $p = 0.8$, and constant firing times. **(c)** Same as (a), but with stochastic firing times, with a standard deviation of 0.1. **(d)** Same as (b), but with stochastic firing times, with a standard deviation of 0.1.

make predictions about real organisms, details such as the values of the parameters predicted by the model for the transition from the clustering to the isolated configuration should be taken with a grain of salt; this is a very idealised model of DNA replication, after all. However, its overall predictions are robust and do not depend strongly on the details of the model. In the Introduction we argue that, all other factors being equal, a shorter S-phase is advantageous, and would be favoured by natural selection. This implies that we should expect to see a significant statistical trend in favour of origin configurations leading to shorter replication times. This leads us to two main predictions:

(a) The typical distance between neighbouring origins with high competence should be greater than the distance between low-competence origins;

(b) The typical distance between neighbouring origins with large firing time differences should be shorter than the distance between origins with small firing time differences.

Notice that we do not expect *all* pairs of low-competence origins to be very close to each other: many factors other than the replication time are expected to affect the evolution of origin placement, and it would be unrealistic to expect the correlations found in real data to match perfectly those predicted by our model. However, we do expect to see statistically significant trends in the origin distribution in the direction indicated by these two predictions.

III. COMPARING MODEL PREDICTIONS WITH EXPERIMENTAL DATA

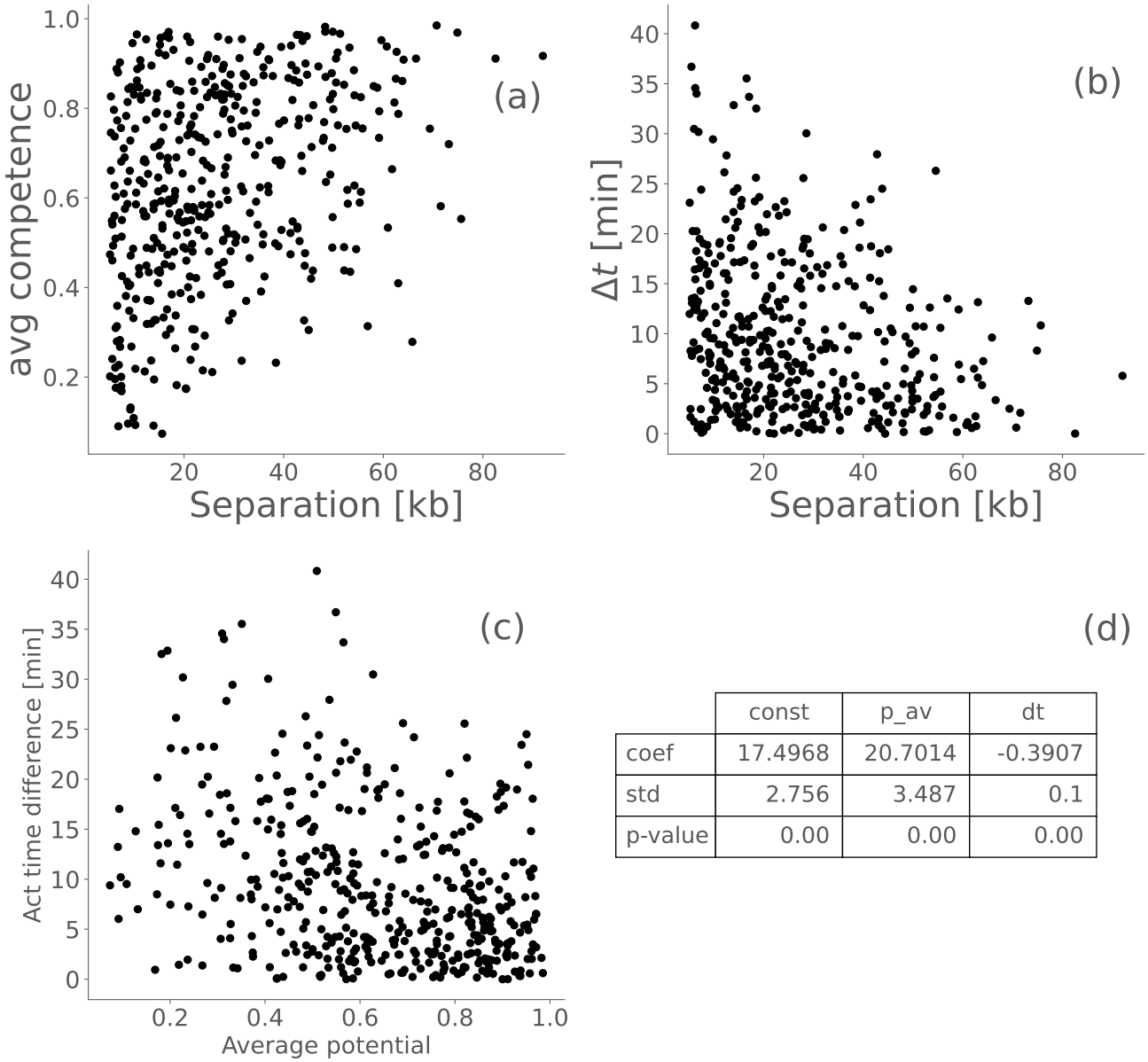
We used two genome-wide origin data sets in order to compare the two main predictions of the model to available data. The first is reported in [20], and consists of the positions of all origins in *S. cerevisiae*, along with a number of origin parameters obtained by fitting next-generation sequencing replication profiles [19] to a whole-genome mathematical model of replication. For our purposes, only the location, competence and mean replication times of each origin are relevant. The second data set is taken from [23]. This data does not include competence data, but it does include firing time data. It also differs from the data reported in [20] in that it uses a slightly different set of origins for its fitting.

A. Neighbouring origins with low competence tend to cluster together

In Fig. 4a, we plot the distances between pairs of neighbouring origins versus the average of the competence of the two origins in each pair. Each point in Fig. 4a represents a pair of neighbouring origins, with origin data taken from all chromosomes; in total, 443 pairs are plotted, resulting from 459 origins distributed in the 16 chromosomes of *S. cerevisiae*. One prediction from our model is that there should be a bias in the distribution of origin placement favouring short distances for pairs of low-competence origins, and larger distances for pairs with high-competence origins. This correlation is indeed seen in the figure: there is a clear trend towards a greater horizontal spread of points for greater competences, establishing that the average distance between origins grows with the average competence.

Another way to see this correlation is by classifying the nearest-neighbour pairs into three categories: the *high-high* pairs consist of pairs where both origins have competences greater than some threshold value p_0; the *low-low* pairs are those where both origins have competences below p_0; and the *high-low* pairs are the ones with a low- and a high-competence origin. In total, there are 173 high-high pairs, 78 low-low pairs, and 192 high-low pairs.

In Fig. 5, we plot histograms of the distribution of distances between nearest-neighbouring origins for each of the three categories listed above, with the threshold value chosen to be $p_0 = 0.6$ (that is, origins are considered "high" if their competence is at least 60%). Fig. 5 confirms that the spread in the distribution of distances for the high-high origin pairs is much greater than for low-low cases. It is clear from the figure that we are much



	const	p_av	dt
coef	17.4968	20.7014	-0.3907
std	2.756	3.487	0.1
p-value	0.00	0.00	0.00

FIG. 4. In (a), (b) and (c), each point represents one pair of neighbouring origins in *S. cerevisiae*. In (a), the horizontal axis is the distance between the origins in each pair, and the vertical axis is the average of their competence. In (b), tho horizontal axis is as in (a), and the vertical axis is the difference in average activation time between the two origins. (c) shows the mean competences versus activation time differences for each pair. Finally, (d) shows the results of multi-variable linear regression for the inter-origin distances, using the activation time differences (**dt** on the table) and mean competences (**p_av**) as predictor variables. For each predictor variable, the value of the fitted coefficient, its standard deviation, and the corresponding p-value are listed. The data is taken from [20].

more likely to see large inter-origin separations for the high-high case. In fact, comparing Figs. 5a and 5b, we see that only a very small fraction of pairs have distances greater than 30 kilobases in the low-low case, whereas pairs with distances greater than 40 kilobases are quite common for high-high pairs.

The details of Fig. 5 of course depend on the precise value of p_0, but for any choice of p_0 greater than 0.5, we see a marked difference between the distance distributions for the high-high compared to the low-low pairs, always with the high-high case favouring greater distances between origins.

One may raise the question of whether the correlations between the origin parameters might be a numerical arti-

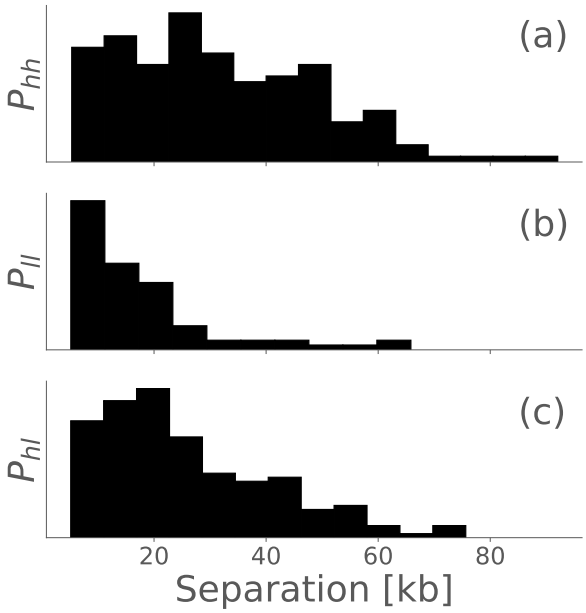


FIG. 5. Histograms of the distance between pairs of neighbouring origins, according to their competence. (a) shows data for origin pairs where both origins have high competence (high-high pairs); (b) plots the case where both origins have low competence (low-low pairs); and (c) is the case where one origin has high competence and the other has low competence. For the purposes of this plot, origins with competence greater than 0.6 are considered to have high competence. The values of the frequencies in the vertical axes are not relevant, and are therefore omitted. The origin competence data was taken from [20].

fact from the fitting process. In order to investigate this, we have simulated replication profiles of populations of artificial virtual chromosomes, with origins whose parameters and positions were all chosen randomly, and then applied the same fitting procedure used to estimate the origin parameters in [20]. The details are in the Supplementary Material. The conclusion is that the fitting of next-generation sequence data does **not** create spurious correlations between origin parameters and their positions that are strong enough to affect our conclusions. We can thus be confident that the correlations shown in 4 and 5 are real.

Figures 4 and 5 provide strong evidence that origin competences and inter-origin distances are correlated in the way predicted by our theory. This is confirmed by calculating the corresponding p-value, which turns out to be 6×10^{-13}. This indicates that the correlation is real, and not a product of blind luck. We also use a more sophisticated Bayesian approach to test the significance of the correlation. This is done in the Supplementary Material. The result again confirms that the correlations

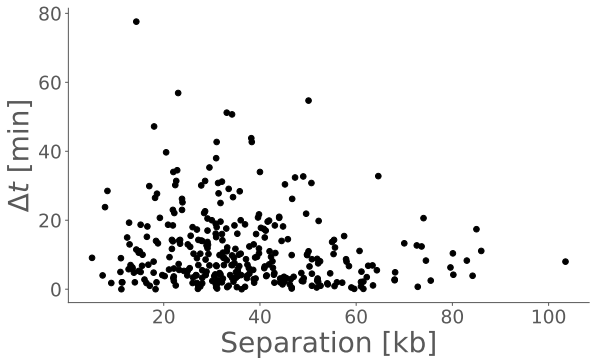


FIG. 6. Each point represents one pair of neighbouring origins in *S. cerevisiae*. The horizontal axis is the distance between the origins in both plots. The vertical axis is the difference in average activation time between the two origins. The data was taken from [23].

are real.

We conclude that there is enough evidence of a strong correlation between origin competences and inter-origin distance, and that this correlation is as predicted by our theory.

B. Neighbouring origins with higher differences in firing times are more likely to be close to each other

Our theory also predicts correlations between origin activation times and their distances to each other. In Fig. 4b, we plot the inter-origin distance versus the absolute value of the difference in firing times for each nearest-neighbouring origin pair in *S. cerevisiae*, with data taken from [20].

It is a clear from Fig. 4b that neighbouring origins with greater differences in origin firing time tend to be closer to each other, as we predict. The p-value of this data is 10^{-8}, confirming that the correlation is not a product of chance. We again apply Bayesian inference to assess more rigorously how strongly the origin data supports our hypothesis. The conclusion is again that the correlation is real; for details, see the Supplementary Material.

There is some evidence that the competences and the firing times are correlated [20]: more competent origins tend to fire earlier. This can be seen in Fig. 4c. It is therefore a legitimate question whether the trent in firing time differences we see in Fig. 4b is simply a consequence of the correlation between potentials and inter-origin distances. In order to see if there is statistically significant evidence that the two trends are real, independently of this correlation, we fit a multiple-variable linear regression model, with the inter-origin distances d as the response variable, and the mean competences p_{av} and firing time differences τ as the predictor variables.

In other words, we fit the statistical model below to the origin pair data:

$$d = ap_{av} + b\tau + c + \epsilon, \tag{13}$$

where c is the constant term in the regression, and ϵ represents noise. The results of this fitting are displayed on the table shown in Fig. 4d. The fact that the p-value for both predictor variables is very low shows that there is statistically significant evidence that both trends are real, and neither can be explained away by the correlation between the two variables.

We have also analysed yeast origin data obtained from the replication model proposed by the Rhind and Bechhoefer groups [23, 31]. Their model does not consider origin competence; they assume that all origins have been licensed at the start of S-phase. Another difference between their model and the one in [20] is that they consider a different set of origins in their model, which has large overlap with the one we use, but is not identical to it. They infer the origins' average firing times for their model by fitting the model's predictions to measured replication profiles, and these results are plotted in Fig. 6. The figure shows a negative correlation between the origins' separations and their firing time differences, in accordance with our prediction. The correlation is not as strong as the one we get from the data in [20], but the p-value is still less than 1%.

We conclude that the origin data from both data sets support our prediction that neighbouring origins with higher differences in firing times are more likely to be close to each other than pairs with similar firing times.

IV. DISCUSSION

Our results support the hypothesis that natural selection favours placements of replication origins in *S. cerevisiae* leading to short replication times. This selection pressure is probably due to the greater vulnerability of cells during DNA replication [24, 25]. Even if the differences in replication time between two competing strains are small, over many generations this can result in one strain thriving and the other going extinct: natural selection is an efficient amplifier of small differences [32].

We note that replication time also plays a crucial role in the *random-gap problem*, which has been extensively investigated in organisms where the locations of origins are not fixed, such as the model organism *Xenopus laevis* [33–35].

A tacit assumption in this work is that the origin locations are a flexible parameter for natural selection to act on, and that mutations resulting in changes in their genome locations are common. This is supported by the fact that events such as genome transpositions are very common in the evolutionary history of most Eukaryotes, including yeast, and they usually lead to origin displacements [26, 36].

If origin locations are selected for their short replication times, why does their distribution not follow the predictions of the theory more closely? The trends seen in Fig. 4 are in the direction predicted by our theory, but the sharp transition from clustered to isolated configurations, predicted by the two-origin model, is nowhere to be seen: although origins with high competence are clearly much more likely to be isolated from nearby origins, some of them do have close neighbours. One of the reasons is the simplicity of our replication model. Real chromosomes and origins are much more complex than our idealised versions of them; to cite just one example, yeast chromosomes typically have dozens of origins, not just two. In addition, there are restrictions in the origin positions which we did not account for in our model; for example, origins usually are not found within genes. And finally, another reason for the imperfect match between our theory and the data is the fact that there are myriad selection pressures acting on the genome which also affect the locations of origins, and are completely independent of the replication timing. So when genes are inserted, deleted or duplicated in the normal course of evolution, distances between origins often change, and sometimes new origins are introduced or existing ones are deleted. The end result is that the overall bias predicted by the theory is still visible, but the fine details predicted by the theory — such as the sharp clustered-isolated transition — are lost.

This study focused on *S. cerevisiae* as a model organism, because it is an organism for which the locations, competences and firing times of all replication origins have been determined. However, many unicellular Eukaryotes have replication origins with fixed positions, and we expect our predictions to hold for these organisms. In addition, many aspects of the replication dynamics are conserved across species, including the locations of most high-competence origins [36].

This paper has focused on the minimisation of replication time as an important selection factor driving the distribution of replication origins. There are other important selection pressures acting on origin positions, however. In [37, 38], the authors focus on a different selection criterion, namely the minimisation of the probability that a double fork stalling event prevents a region of the genome from finishing replication [2]. The resulting "best" origin distribution has the origins regularly spaced from one another, like beads on a string. This may seem at odds with our predictions, since we predict that origins will sometimes bunch together and sometimes stay apart, depending on their competences and firing times; but in fact, the prediction in [37] is what we would expect from our model if all origins had high competence. But the majority of origins in yeast have high competences — 66% of the origins have competence greater than 0.5. The results reported on [37] are therefore broadly consistent with our model, because the high-competence origins dominate the replication dynamics of yeast. This implies that both selection pressures on origin placement — the

minimisation of replication time and the minimisation of double fork stalling events — result in broadly similar origin distributions.

V. CONCLUSION

We have shown that the positions of replication origins in yeast are strongly correlated with the parameters determining the dynamics of origin firing, namely the competence and average firing time. Specifically, we found that neighbouring origins with low competence are more likely to be clustered together in the genome; and that neighbouring origins with large differences in their firing times are more likely to be clustered. We showed how these correlations are predicted from a mathematical model incorporating the assumption that the positions of replication origins are biased towards configurations minimising the total replication time. This lends support to our hypothesis that the minimisation of replication time is a selection pressure that plays an important role in shaping the spatial distribution of replication origins in the genome of yeast, and potentially in the genomes of other Eukaryotes as well.

VI. ACKNOWLEDGEMENTS

We want to acknowledge Conrad Nieduszynski for valuable discussions.

For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising from this submission.

[1] N. Agier, S. Delmas, Q. Zhang, A. Fleiss, Y. Jaszczyszyn, E. V. Dijk, C. Thermes, M. Weigt, M. Cosentino-Lagomarsino, and G. Fischer, The evolution of the temporal program of genome replication, Nature Communications **9**, 2199 (2018).

[2] R. Droghetti, N. Agier, G. Fischer, M. Gherardi, and M. C. Lagomarsino, An evolutionary model identifies the main evolutionary biases for the evolution of genome-replication profiles, eLife **10**, e63542 (2021).

[3] S. P. Bell and A. Dutta, Dna replication in eukaryotic cells., Annual review of biochemistry **71**, 333 (2002).

[4] M. Méchali, Eukaryotic dna replication origins: Many choices for appropriate answers, Nature Reviews Molecular Cell Biology **11**, 728 (2010).

[5] M. Méchali, K. Yoshida, P. Coulombe, and P. Pasero, Genetic and epigenetic determinants of dna replication origins, position and activation, Current Opinion in Genetics and Development **23**, 124 (2013).

[6] A. C. Leonard and M. Mechali, Dna replication origins, Cold Spring Harbor Perspectives in Medicine **3**, a010116 (2013).

[7] M. Fragkos, O. Ganier, P. Coulombe, and M. Méchali, Dna replication origin activation in space and time, Nature Reviews Molecular Cell Biology **16**, 360 (2015).

[8] R. C. Alver, G. S. Chadha, and J. J. Blow, The contribution of dormant origins to genome stability: From cell biology to human genetics, DNA Repair **19**, 182 (2014).

[9] M. W. Parker, M. R. Botchan, and J. M. Berger, Mechanisms and regulation of dna replication initiation in eukaryotes, Critical Reviews in Biochemistry and Molecular Biology **52**, 107 (2017).

[10] I. Attali, M. R. Botchan, and J. M. Berger, Structural mechanisms for replicating dna in eukaryotes, Annual Review of Biochemistry **90**, 77 (2021).

[11] Y. C. Lin and S. G. Prasanth, Replication initiation: Implications in genome integrity, DNA Repair **103**, 103131 (2021).

[12] L. Mei and J. G. Cook, Efficiency and equity in origin licensing to ensure complete dna replication, Biochemical Society Transactions **49**, 2133 (2021).

[13] S. Ticau, L. J. Friedman, K. Champasa, I. R. Corrêa, J. Gelles, and S. P. Bell, Mechanism and timing of mcm2–7 ring closure during dna replication origin licensing, Nature Structural and Molecular Biology **24**, 309 (2017).

[14] J. J. Blow and A. Dutta, Preventing re-replication of chromosomal dna, Nature Reviews Molecular Cell Biology **6**, 476 (2005).

[15] E. E. Arias and J. C. Walter, Pcna functions as a molecular platform to trigger cdt1 destruction and prevent re-replication, Nature Cell Biology **8**, 84 (2006).

[16] A. P. S. de Moura, R. Retkute, M. Hawkins, and C. a Nieduszynski, Mathematical modelling of whole chromosome replication, Nucleic acids research **38**, 5623 (2010).

[17] K. Shirahige, T. Iwasaki, M. Rashid, N. Ogasawara, and H. Yoshikawa, Location and characterization of autonomously replicating sequences from chromosome vi of saccharomyces cerevisiae, Molecular and Cellular Biology **13**, 5043 (1993).

[18] M. Yamashita, Y. Hori, T. Shinomiya, C. Obuse, T. Tsurimoto, H. Yoshikawa, and K. Shirahige, The efficiency and timing of initiation of replication of multiple replicons of saccharomyces cerevisiae chromosome vi, Genes to Cells **2**, 655 (1997).

[19] C. A. Müller, M. Hawkins, R. Retkute, S. Malla, R. Wilson, M. J. Blythe, R. Nakato, M. Komata, K. Shirahige, A. de Moura, and C. A. Nieduszynski, The dynamics of genome replication using deep sequencing, Nucleic Acids Research **42**, e3 (2014).

[20] M. Hawkins, R. Retkute, C. a Müller, N. Saner, T. U. Tanaka, A. P. S. de Moura, and C. a Nieduszynski, High-resolution replication profiles define the stochastic nature of genome replication initiation and termination, Cell reports **5**, 1132 (2013).

[21] R. Retkute, C. Nieduszynski, and A. de Moura, Dynamics of dna replication in yeast, Physical Review Letters **107**, 068103 (2011).

[22] R. Retkute, C. Nieduszynski, and A. de Moura, Mathematical modeling of genome replication, Physical Review E **86**, 031916 (2012).

[23] S. C.-H. Yang, N. Rhind, and J. Bechhoefer, Modeling

genome-wide replication kinetics reveals a mechanism for regulation of replication timing, Molecular systems biology **6**, 404 (2010).

[24] D. Branzei and M. Foiani, Regulation of dna repair throughout the cell cycle, Nature Reviews Molecular Cell Biology **9**, 297 (2008).

[25] H. A. Jeiranian, B. J. Schalow, C. T. Courcelle, and J. Courcelle, Fate of the replisome following arrest by uv-induced dna damage in escherichia coli., Proceedings of the National Academy of Sciences of the United States of America **110**, 11421 (2013).

[26] S. C. D. Rienzi, K. C. Lindstrom, T. Mann, W. S. Noble, M. K. Raghuraman, and B. J. Brewer, Maintaining replication origins in the face of genomic change, Genome Research **22**, 1940 (2012).

[27] J. Karschau, J. Blow, and A. de Moura, Optimal placement of origins for dna replication, Physical Review Letters **108**, 058101 (2012).

[28] B. Theulot, L. Lacroix, J. M. Arbona, G. A. Millot, E. Jean, C. Cruaud, J. Pellet, F. Proux, M. Hennion, S. Engelen, A. Lemainque, B. Audit, O. Hyrien, and B. L. Tallec, Genome-wide mapping of individual replication fork velocities using nanopore sequencing, Nature Communications **13**, 3295 (2022).

[29] J. Bechhoefer and N. Rhind, Replication timing and its emergence from stochastic processes, Trends in genetics : TIG **28**, 374 (2012).

[30] a. Baker and J. Bechhoefer, Inferring the spatiotemporal dna replication program from noisy data, Physical Review E **89**, 032703 (2014).

[31] S. C.-H. Yang, M. G. Gauthier, and J. Bechhoefer, Computational methods to study kinetics of dna replication, Methods in molecular biology (Clifton, N.J.) **521**, 555 (2009).

[32] J. M. Smith, *Evolutionary genetics*, 2nd ed. (Oxford University Press, 1998).

[33] J. J. Blow, P. J. Gillespie, D. Francis, and D. A. Jackson, Replication origins in xenopusegg extract are 5–15 kilobases apart and are activated in clusters that fire at different times, The Journal of Cell Biology **152**, 15 (2001).

[34] M. G. Gauthier and J. Bechhoefer, Control of dna replication by anomalous reaction-diffusion kinetics, Physical review letters **102**, 158104 (2009).

[35] A. Baker, B. Audit, S. Yang, J. Bechhoefer, and A. Arneodo, Inferring where and when replication initiates from genome-wide replication timing data, Physical Review Letters **108**, 268101 (2012).

[36] C. A. Müller and C. A. Nieduszynski, Conservation of replication timing reveals global and local regulation of replication origin activity, Genome research **22**, 1953 (2012).

[37] T. J. Newman, M. a Mamun, C. a Nieduszynski, and J. J. Blow, Replisome stall events have shaped the distribution of replication origins in the genomes of yeasts, Nucleic acids research **41**, 9705 (2013).

[38] M. A. Mamun, L. Albergante, A. Moreno, J. T. Carrington, J. J. Blow, and T. J. Newman, Inevitability and containment of replication errors for eukaryotic genome lengths spanning megabase to gigabase, Proceedings of the National Academy of Sciences of the United States of America **113**, E5765 (2016).