PHYSICAL REVIEW E **86**, 031916 (2012)

Mathematical modeling of genome replication

Renata Retkute and Conrad A. Nieduszynski
Centre for Genetics and Genomics, University of Nottingham, Nottingham NG7 2UH, United Kingdom

Alessandro de Moura
Institute of Complex Systems and Mathematical Biology, University of Aberdeen, Aberdeen AB24 3UE, United Kingdom

(Received 8 February 2012; revised manuscript received 15 August 2012; published 17 September 2012)

Eukaryotic DNA replication is initiated from multiple sites on the chromosome, but little is known about the global and local regulation of replication. We present a mathematical model for the spatial dynamics of DNA replication, which offers insight into the kinetics of replication in different types of organisms. Most biological experiments involve average quantities over large cell populations (typically $>10^7$ cells) and therefore can mask the cell-to-cell variability present in the system. Although the model is formulated in terms of a population of cells, using mathematical analysis we show that one can obtain signatures of stochasticity in individual cells from averaged quantities. This work generalizes the result by Retkute *et al*. [Phys. Rev. Lett. **107**, 068103 (2011)] to a broader set of parameter regimes.

DOI: 10.1103/PhysRevE.86.031916 PACS number(s): 87.14.gk, 87.10.Ca, 87.10.Mn

I. INTRODUCTION

DNA replication, the process during which cells' genetic information is duplicated, is one of the most fundamental processes in biology. Eukaryotic cells regulate the replication of their genomes in a highly complex manner: it is vital that chromosomal replication be completed before cell division takes place, in order to pass full and accurate genetic information to the daughter cells.

Cell cycle progression in eukaryotic organisms consists of four morphologically distinct phases: a gap phase (G1) during which a cell grows, followed by the DNA synthesis (S) phase, when the cell's genome is duplicated, a second gap phase (G2), and the mitosis (M) phase, when the cell divides into two [1]. The duration of each of these phases varies from organism to organism, but in all eukaryotes chromosome replication occurs during the S phase and is initiated at specific locations on the chromosome called *replication origins*. When an origin is activated, two replication forks are formed, which travel on the chromosome in opposite directions. The number of origins varies depending on species and cell type; for example, most bacterial genomes are replicated from a single origin. The size of eukaryotic genomes necessitates the use of multiple origins to ensure timely and complete replication prior to cell division. The use of multiple origins requires tight regulation of origin activity to ensure that sufficiently many origins are activated, but that no origin is activated more than once in a single round of genome replication. This is achieved by the mechanism of *licensing* of replication origins. This consists of binding of a series of specific protein complexes at origin sites in the DNA, followed by the loading of pairs of Mcm2-7 molecules. If in a given cell licensing of a certain origin is not completed by the time the S phase starts, that origin is unable to function [2]. Also regulated are the number of origins that are activated in a given S phase and timing of the replication of specific regions on a chromosome [3].

There has been much interest recently in the mathematical modeling of DNA replication [4–15]. Two different modeling approaches have been used: simulations to capture the replication dynamics at a single cell level [8–12], and probabilistic models that characterize the dynamics of replication at a population level [5,7,13–15]. Although valuable insights have been gained from previous works on mathematical modeling, they ignore the possibility that origins can fail to license, and we will show that this has a crucial effect on the system's dynamics [4,5]. In addition, most of the existing models are numerical.

In this work, we analyze an analytical model of eukaryotic DNA replication which fully takes into account the stochastic nature of both the licensing process and origin activation [5]. We start by formulating a general model for organisms with multiple chromosomes and multiple origins; then we analyze in detail the kinetics of replication for an idealized chromosome with two origins, where the most important features of the replication dynamics can be studied in its simplest nontrivial case. We focus on quantities which are of biological interest: replication profiles, average number of active origins, replicon sizes, and others; and we derive explicit analytical expressions for them.

Genome replication has been comprehensively studied in the model organism *Saccharomyces cerevisiae* (brewer's yeast). High-throughput experiments have allowed the measurement of replication times as a function of chromosomal position for the whole genome [16]. These methods yield average replication times at distinct chromosomal positions, over large cell populations (typically $>10^7$ cells), and therefore can mask the cell-to-cell variability present in the system [17]. To date single cell and single molecule studies are not able to measure the kinetics of whole genome replication [17,18]. The low abundance of the molecules involved in triggering origin activation strongly suggests that origins have stochastic activation times [19], with origins having different relative activation probabilities [20].

Replication origins in eukaryotic organisms must successfully complete the process of licensing in order to be able to replicate during the S phase. Licensing involves a number of orchestrated binding events between origin sites and molecules, some of which are present in low numbers in the cell. Once the S phase starts, no further licensing is allowed, so each origin has a relatively short time window in which to complete licensing. This suggests that a given origin will not manage to complete licensing in every cell within a population before the onset of the S phase, due to inhomogeneities in the abundance of key molecules caused by stochasticity in their expression. We anticipate that different origins will have different licensing efficiencies, due to known differences in affinities to the various species involved in licensing [21], as well as stochasticity in the chemical dynamics of licensing. This leads us to the concept of *origin competence* [4], which is the probability that a given origin in a population will finish the licensing process before the S phase, and is thus eligible for originating replication forks. The competence of an origin is hard to measure directly, although plasmid replication efficiency experiments suggest that many origins in yeast have low competence, making the concept relevant to the study of the dynamics of replication. The concept of competence introduces an additional parameter to each origin in a chromosome, which has the technical disadvantage that it makes parameter estimation from data somewhat problematic, as discussed in [4]. Despite this, we feel there are compelling biological reasons to take this effect into account and investigate its consequence for DNA replication dynamics; this is one of the main goals of this work.

We note that this idea of origins with less than 100% competence is compatible with the hypothesis that origins have a probability distribution of having different numbers of Mcm2-7 molecules, which is proposed in [20] to determine their activation times. But any mechanism generating a stochastic distribution of Mcm2-7 numbers in an origin will have a nonzero probability of not having a pair of Mcm2-7 molecules in any particular origin; this corresponds exactly to the origin having failed to license, and therefore having a competence below 100%.

We notice that the dynamics of DNA replication has many similarities with the process of nucleation, and some models of DNA replication [7,14] are closely related to Kolmogorov's classical model of nucleation [22]. The model we propose here can be regarded as an inhomogeneous model of nucleation with quenched disorder, where nucleation starts at specific sites. Inhomogeneous models of nucleation have been studied in the context of statistical physics and have relevance to surface science and other areas [23,24].

II. THE MODEL

In our model we consider a chromosome with N origins, where each origin i is defined by the following parameters: its chromosomal position x_i; the probability q_i that the origin achieves licensing (in a given cell within a population), and is thus capable of activating, i.e., competence; and the activation time probability distribution $p_i(t)$, which is the probability density of origin i activating and starting bidirectional replication forks at time t [4]. Since an origin may not be competent in every cell within the population, in general $q_i < 1$, and p_i satisfies

$$\int_{-\infty}^{+\infty} p_i(t)dt = q_i.$$

The fundamental quantity from which all statistical properties of this system can be calculated is the probability density $P(x,t)$, defined such that $P(x,t)dt$ is the probability that chromosomal position x is replicated between times t and $t+dt$. If only origin i were present, P would be given by $P(x,t) = p_i(t - |x - x_i|/v)$, where v is the fork velocity, which we assume to be a constant.

In the presence of all N origins, the calculation of $P(x,t)$ is complicated by the fact that position x can be replicated by forks originated from any of the origins [4,14]. Let us assume that position x is replicated between times t and $t+dt$ by a fork from origin i. This requires that (i) origin i activated at time $t - |x - x_i|/v$, so that the fork arrives at x at time t; and (ii) all other origins $j \neq i$ either have not activated or they have activated but their forks would arrive at x later than t. This allows us to account for *passive* replication—i.e., an origin is inactive due to replication by a fork that originated at another origin The probability density for event (i) is

$$p_i(x,t) = p_i(t - |x - x_i|/v), \tag{1}$$

and the probability for event (ii) is

$$Q_i(x,t) = \prod_{j \neq i} M_j(x,t), \tag{2}$$

where M_i is the probability that a fork from origin i arrives later than t, or fails to activate:

$$M_i(x,t) = s_i + \int_t^{+\infty} p_i(x,\tau)d\tau, \tag{3}$$

where $s_i = 1 - q_i$ is the probability of origin i not being competent. Therefore, the probability density $P_i(x,t)$ that position x is replicated by origin i at time t is

$$P_i(x,t) = p_i(x,t)Q_i(x,t). \tag{4}$$

Finally, the probability density that position x is replicated at time t, irrespective of which origin the fork started from, is [5]

$$P(x,t) = \sum_{i=1}^{N} P_i(x,t). \tag{5}$$

Using Eq. (5), expressions for various quantities of biological interest can be found. We notice here that similar expressions have been derived before for the case where all the origins are 100% competent [14].

III. GENERAL RESULTS FOR THE MODEL

Now we will derive expressions for important quantities characterizing the replication dynamics, which are of great interest to biologists working in this area: the mean replication time; the efficiency of origins, that is, in what fraction of cells a given origin has activated in a round of replication; the fraction f of total DNA replicated (or the *relative DNA content*) at any given time in the S phase; the rate of origin initiation in a population; the number of replication forks and

active origins as a function of time; the probability of fork termination and how it depends on the chromosomal position; and the interorigin distance distribution between active origins. All these quantities are derived below from the fundamental expression Eq. (5).

A. Mean replication time

One of the most important quantities in the area of DNA replication biology is the *replication profile*, which is the mean replication time $T(x)$ at the chromosomal position x; here the average is over a large population of cells, which is the typical situation in experiments. There are a number of techniques for obtaining whole-genome replication profiles, such as the density transfer method or by measuring relative DNA copy number using microarray [16,25] or next generation sequencing [26,27] techniques. In the biological literature, the term "replication profile" has a number of different but related meanings: mean replication time, mean percentage replicated, mean copy number at a chromosomal position; or the so-called S:G1 ratio, where cells in the G1 and S phases of the cell cycle are sorted and numbers of cells in both phases are compared for each sequence [27–33]. All of those quantities are closely related to the mean replication time $T(x)$, and they contain essentially the same information. Therefore in this paper the term "replication profile" will mean $T(x)$.

From Eq. (5), we can write the mean replication time as

$$T(x) = \frac{1}{1-\prod_{i=1}^{n} s_i} \int_{-\infty}^{+\infty} t P(x,t) dt, \tag{6}$$

where the normalization factor $1-\prod_{i=1}^{n} s_i = \int_{-\infty}^{+\infty} P(x,t)dt$ is the probability that at least one of the origins will activate. In Eq. (6) we are thus excluding the situation where all origins simultaneously fail from the definition of the average; the probability of this happening is very remote in real cells. The probabilities will be defined in this way in all the remaining expressions in this paper.

The replication profile ([$T(x)$ curves] has been measured in a number of organisms. However, caution is required when interpreting $T(x)$ curves. In some of the biological literature $T(x)$ curves are used to directly infer origin parameters [16]. For example, it is widely accepted that the values of T at x_i are the average activation times of origins. However, Eq. (6) shows that $T(x)$ is determined collectively by all origins [4,5], which suggests that simple interpretations of $T(x)$ are likely to be misleading. This issue will be discussed in more detail in Sec. VII.

B. Fraction of DNA replicated

The fraction of cells in a population which have position x on the chromosome replicated by time t can be calculated by integrating the probability $P(x,t)$ with respect to t [5]:

$$m(x,t) = \frac{1}{1-\prod_{i=1}^{n} s_i} \int_{-\infty}^{t} P(x,\tau) d\tau. \tag{7}$$

The higher the value of the fraction replicated, the earlier the chromosome position replicates on average during the S phase [16].

Replication dynamics has also been studied by measuring the *copy number* change by differentiating between replicated and nonreplicated stages of the cell cycle [25,27]. The copy number $C(x,t)$ of the position on the genome at a particular time t is given by

$$C(x,t) = m(x,t) + 1. \tag{8}$$

A twofold increase in copy number is required as the cell progresses from the G1 to the G2 phase.

A relationship between mean copy number $\bar{C}(x)$ and mean replication time $T(x)$ for the S phase which starts at time T_1 and finishes at time T_2:

$$\bar{C}(x) = \frac{T_2 - T(x)}{T_2 - T_1} + 1; \tag{9}$$

for details see Appendix A.

The *replicated fraction* $f(t)$ of a genome of length L is given by the average over the whole length of the chromosome of the probability that a piece of DNA has replicated by time t:

$$f(t) = \frac{1}{L} \int_0^L m(x,t) dx. \tag{10}$$

As the population of cells enters into the S phase and replication starts, the chromosomal content is equal to one copy of the genome, and the replicated fraction $f(t)$ is 0. As replication progresses, $f(t)$ increases and towards the end of the S phase the DNA content doubles, ensuring a full complement of DNA for each daughter cell. By measuring the fraction of replicated DNA as a function of time, it is possible to determine the rate of total DNA synthesis during the cell cycle [34].

Figure 1(a) shows the simulated replicated fraction for a virtual chromosome with a length 100 kilobase pairs (kb) with ten origins periodically positioned along a chromosome. In this example, each origin has a Gaussian activation distribution with a mean of 15 min and a standard deviation of 5 min. We see that the fraction of DNA replicated as a function of time has a sigmoidal shape.

C. Rate of origin initiation

The probability density of a single isolated origin i activating and starting bidirectional replication forks at time t is $p_i(t)$ (Sec. II). Any given origin i can activate only if forks from neighbor origins $j \neq i$ arrive at position x_i later than time t, or fail to activate, and this probability is given by $\prod_{j\neq i} M_j(x_i,t)$, with M_j defined by Eq. (3). Then the rate of origin initiation $g(t)$ is defined as the sum over all origins of the probability densities of the origins activating at a time t:

$$g(t) = \frac{1}{1-\prod_{i=1}^{n} s_i} \sum_{i=1}^{N} p_i(t) \prod_{j\neq i} M_j(x_i,t). \tag{11}$$

The importance of this quantity lies in the fact that in a large population of cells, the number of origins activating in a small time window $(t, t+dt)$ is proportional to $g(t)$; hence the term "rate" used above.

Population-averaged measurements of replication initiations per time unit per unit length of *unreplicated* DNA

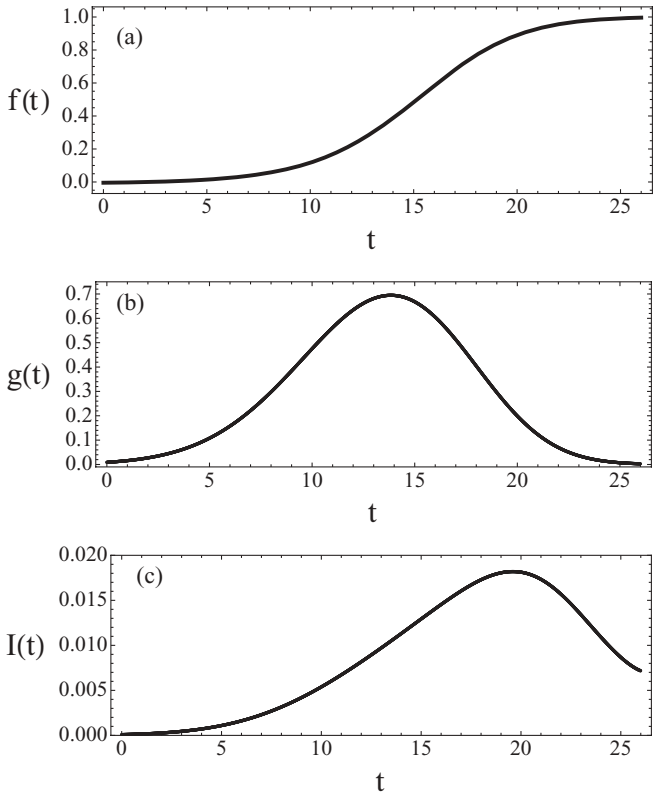

FIG. 1. Results for a virtual chromosome of length 100 kb with ten origins periodically positioned along the chromosome. Origin activation follows a Gaussian distribution with a mean of 15 min, a standard deviation of 5 min, and a fork velocity 1.5 kb/min. (a) Fraction of replicated DNA $f(t)$, (b) rate of origin initiation $g(t)$, and (c) initiation rate $I(t)$.

have been obtained for a number of organisms, including *S. cerevisiae*, *Schizosaccharomyces pombe*, *Drosophila melanogaster*, *Xenopus laevis*, and *Homo sapiens* [34,35]. All profiles have been observed to possess a strikingly similar shape: increasing during the first half of the S phase and then decreasing before the end of the S phase.

The initiation rate, i.e., the rate of origin initiations per time unit per unit length of unreplicated DNA, is obtained by dividing Eq. (11) by the chromosome length L and by the fraction of DNA still unreplicated at time t [obtained from Eq. (10)]:

$$I(t) = \frac{g(t)}{L[1 - f(t)]}. \tag{12}$$

To test if the model reproduces the behavior of the initiation rate observed in the experiments described in [35], we have applied Eqs. (11) and (12) to the virtual chromosome described in Sec. III B. We consider a virtual chromosome with a length of 100 kb replicated from ten origins periodically positioned along the chromosome; each origin has a Gaussian activation distribution function with a mean of 15 min and a standard deviation of 5 min. The general shape of the curve predicted for the initiation rate per time unit per unreplicated unit of DNA [$I(t)$; Fig. 1(c)] is in good agreement with the experimental results [35]. In the case we have analyzed, the initiation rate $I(t)$ increases during the first 4/5 of the S phase and then declines for the last 1/5 of the S phase. This late decrease is due to the rate of the origin initiation function $g(t)$ reaching its maximum just after mid-S-phase (at 14 min). The end of the S phase is the time at which the fraction of DNA replicated, $f(t)$, reaches 1 (at 25 min).

D. Origin efficiency and average number of active origins

The efficiency of an origin i is the fraction of cells in a population that actually activated that origin in any given S phase [36]. In any given cell, not all origins will be activated during the S phase; for example, in some mammalian cells most origins are used in less than 10% of cell cycles [26].

The efficiency of origin i is the probability that origin i activates at any time, and is therefore given by the integral of the probability that the ith origin initiates between times t and $t + dt$ multiplied by the probability that none of the other origins has replicated position x_i before time t [4,14]:

$$e_i = \frac{1}{1 - \prod_{i=1}^{N} s_i} \int_{-\infty}^{\infty} p_i(t) \prod_{j \neq i} M_j(x_i,t) dt. \tag{13}$$

It has been observed in experiments that the efficiency of origins depends heavily on the distance and timing of activation of neighboring origins [37]. This is seen directly from Eq. (13), where the term $\prod_{j \neq i} M_j(x_i,t)$ encodes the influence of other origins on the chromosome.

Another quantity describing the collective dynamics of DNA replication is the average number n_O of active origins, that is, on average how many origins have activated in a cell within a population. The presence of multiple origins creates redundancy and the average number of activated origins will be less than the total number of origins present on a chromosome.

The average number of active origins is readily obtained from the origin initiation rate given by Eq. (11):

$$n_O = \int_{-\infty}^{\infty} g(t) dt. \tag{14}$$

Applying Eq. (11) to Eq. (14) and using Eq. (13), we can relate the average number of active origins to origin efficiencies:

$$n_O = \sum_{i=1}^{n} e_i \tag{15}$$

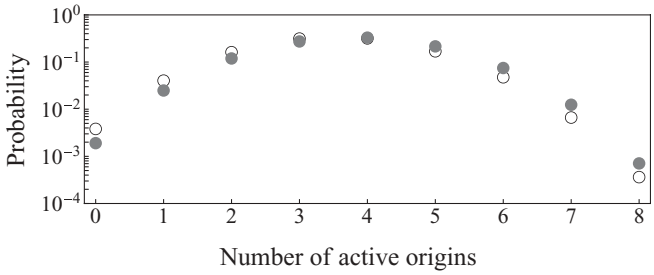

FIG. 2. Probability distribution for a number of activated origins on *S. cerevisae* chromosome VI: experimental results from [19] (black full circles) and calculations based on parameter values from [4] (gray empty circles).

Figure 2 shows the probability distribution for a number of activated origins on *S. cerevisae* chromosome VI, based on origin efficiencies determined experimentally [19] and calculations based on normal activation distribution and parameter values from [4]. The average number of active origins n_O on chromosome VI, calculated using Eq. (15) based on experimental data and parameter estimation, gave 3.8 and 3.6 origins per cell cycle, respectively.

E. Describing the dynamics of replication forks

As replication progresses through the S phase, replication forks originate at activated origins, and disappear when two forks moving in opposite directions in the chromosome collide. The average number of forks moving at some time t is a quantity that can yield valuable insights into the replication dynamics. If there were only a single fork moving in the chromosome, the rate $df(t)/dt$ of replication would be simply v, the fork velocity; n_f forks thus correspond to a replication rate of $n_f v$, which leads to the expression

$$n_f(t) = \frac{1}{v}\frac{df(t)}{dt},$$

where $f(t)$ is the fraction of replicated DNA at time t. This expression was also derived for DNA replication in *X. laevis* [38]. By applying Eqs. (7) and (10) in above equation we have

$$n_f(t) = \frac{1}{v\left(1-\prod_{i=1}^{n} s_i\right)}\int_0^L P(x,t)dx, \tag{16}$$

where L is the length of the chromosome.

The direction of replication fork movement can be analyzed using a distribution for the proportion of left moving forks at each position x:

$$n_{\text{left}}(x) = \frac{1}{1-\prod_{i=1}^{n} s_i}\int_{-\infty}^{\infty}\sum_{i=1}^{N}\mathcal{I}(x<x_i)p_i(x,t) \times \prod_{j\neq i} M_j(x,t)dt, \tag{17}$$

where $\mathcal{I}(X)$ is the indicator function, which takes a value equal to 1 if the condition X is satisfied and 0 otherwise.

Figure 3(a) shows the probability distribution for the proportion of leftward traveling forks at position x for *S. cerevisae* (chromosome VI), computed using Eq. (17), with origin parameters estimated in [4]. All regions except for the start and the end of the chromosome are replicated by both leftward and rightward moving forks.

The difference between the proportion of right and left moving forks gives the average fork polarity $f_{pl}(x)$. The average fork polarity is the product of the derivative of the mean replication time and the replication fork velocity [15]:

$$f_{pl}(x) = vT'(x).$$

Taking into account that the sum of the proportion of left and right moving forks is equal to 1, this gives the relationship between the proportion of left moving forks and mean replication time:

$$n_{\text{left}}(x) = \frac{1-vT'(x)}{2}.$$

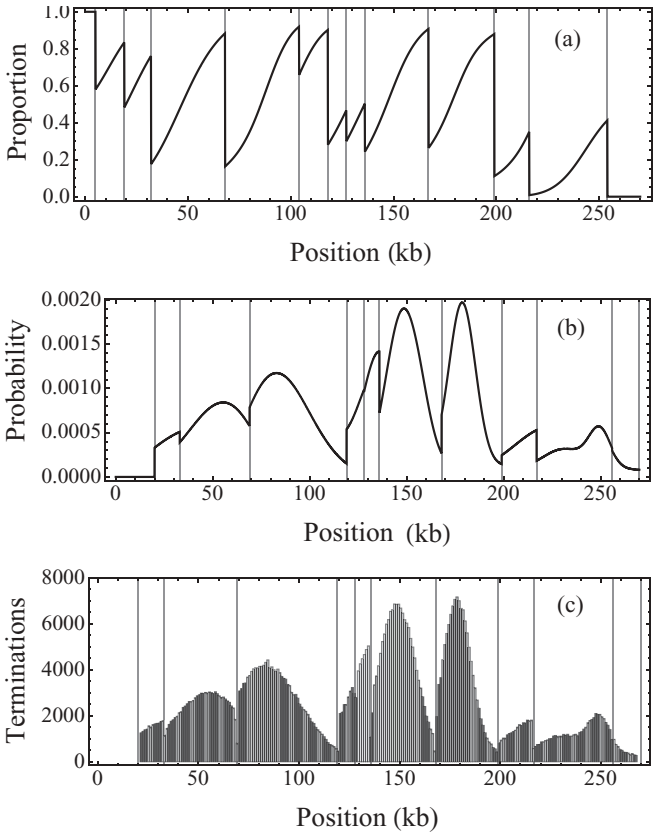

FIG. 3. Probability distribution for (a) the proportion of left moving forks, (b) fork termination at position x on *S. cerevisae* chromosome VI based on Eq. (16), and (c) histogram of fork termination sites based on 10^6 Monte Carlo simulations. Parameter values from [4].

When two forks collide, they both terminate and replisome proteins are released from the DNA molecule. In *Escherichia coli* termination sites are regulated and fixed, but in eukaryotes termination is less well defined than in bacteria, and specific sites are rare [39].

To find the probability density of termination events at position x and at time t, we first consider any given pair of origins i and j with $i<j$—so their positions satisfy $x_i<x_j$. For the forks originating at i and j to terminate exactly at position x at a time t, the following conditions must be met: (a) x must lie between x_i and x_j; (b) i and j must have activated at times $t-\frac{|x-x_i|}{v}$ and $t-\frac{|x-x_j|}{v}$, respectively; and (c) the position x must not be replicated by any other origin k, with $k\neq i, k\neq j$, at time t. The total probability density of fork termination at position x and time t is the sum of the termination probabilities for all possible pairs (i,j) satisfying $x_i\leqslant x\leqslant x_j$; since $i<j$, it is enough to consider indices in the sum given by $i=1,\ldots,N$ and $j=i+1,\ldots,N-1$. Finally, the probability (density) that forks will terminate at x is found by integrating over all times, yielding the

expression

$$P_{\rm ft}(x) = \frac{1}{\mathcal{N}} \int_{-\infty}^{\infty} \sum_{i=1}^{N} \sum_{j=i+1}^{N-1} \mathcal{I}(x_i \leqslant x \leqslant x_j) \times p_i(x,t) p_j(x,t) \prod_{k\neq i, k\neq j} M_k(x,t) dt, \quad (18)$$

where $\mathcal{N}$ is a normalization constant.

Figure 3(b) shows the probability distribution of fork termination at position x for *S. cerevisae* (chromosome VI), computed using Eq. (18), with origins activating according to a normal distribution with mean and variance values estimated in [4]. For a comparison, we have plotted a histogram of fork termination sites based on 10^6 Monte Carlo simulations [Fig. 3(c)], the solution based on Eq. (18) agrees with the simulation results within statistical limits.

F. Distribution of distances between active origins

The distribution of distances between active origins (between origins that have actually activated) in a population of cells is an important quantity for biology. Its importance is related to the fact that the stochastic nature of origin activation can result in large distances between active origins, with the consequence that sections of the genome would take too long to replicate; this can result in portions of the genome remaining unreplicated when cells enter the M phase [40] or causing instability in fragile sites [41].

For organisms such as yeast where the replication origins have fixed positions in any given chromosome, the inter-active-origin distance can assume only a discrete set of values. For example, about 120 early inter-active-origin distances in fission yeast and between 250 and 300 in budding yeast [42] have been observed to have a typical spacing of between 30 and 100 kb [43]. In normal human primary keratinocyte cells, the mean value for inter-active-origin distances is about 124 kb [44].

The distance y_{ij} between origins i and j is equal to $y_{ij} = |x_i - x_j|$, where x_i is the position of origin i and x_j is the position of origin j. Our goal is to obtain an expression for the probability distribution of distances between active origins $\mathcal{P}(y)$ for the interorigin distance y. Consider first a given origin pair (i,j), with $i < j$, then the probability density that this particular pair activates in one replication round, and no other origin activates in between them—this event contributes to $\mathcal{P}$ with y_{ij}. The probability density that origin i activated at time t and neighboring origins on the left either have not activated or they have activated but their forks would arrive at x_i later than t, is given by $p_i(t)\prod_{k=1}^{i-1} M_k(x_i,t)$. The probability density that origin j has activated at time t and neighboring origins on the right either have not activated or they have activated but their forks would arrive at x_j later than t; this is given by $p_j(t)\prod_{k=j+1}^{N} M_k(x_j,t)$. Since $y_{ij} = y_{ji}$ it is sufficient to consider origin pairs with indices i and j with ranges given by $i = 1,\ldots,N$ and $j = i+1,\ldots,N-1$. Further, if i and j are such that $j - i > 1$, i.e., these origins are not adjacent, we need to ensure that all origins k lying between them $(i < k < j)$ will not activate at all (otherwise there would be an origin activating between i and j, and this would split the interval $[x_i,x_j]$ into two); this probability is equal to $\prod_{k=i+1}^{j-1} M_k(x_k,t)$. The probability density function for inter-active-origin distances is given by the integral over all times of the product of these probabilities:

$$\mathcal{P}(y) = \frac{1}{\mathcal{N}} \int_{-\infty}^{\infty} \sum_{i=1}^{N} \sum_{j=i+1}^{N-1} \mathcal{I}(y = |x_i - x_j|) \times p_i(t) p_j(t) \prod_{k=1}^{i-1} M_k(x_i,t) \prod_{k=j+1}^{N} M_k(x_j,t) \times \prod_{k=i+1}^{j-1} M_k(x_k,t) dt, \quad (19)$$

where $\mathcal{N}$ is a normalization constant. The probability density function $\mathcal{P}$ takes nonzero values for y equal to distances separating any two origins on the chromosome.

Figures 4(a) and 4(b) show the probability distribution of distances between active origins for *S. cerevisae* (chromosome VI), computed using Eq. (19) and a histogram for 10^5 Monte Carlo simulation, with normal activation distribution and parameter values estimated in [4]. Based on this distribution, we calculate that the mean distance between active origins is 55 kb. Whole-genome analysis has shown that the average distance between active replication origins in the *S. cerevisiae* genome is approximately 58 kb [45]. We note that the distribution of inter-active-origin distances has a long tail; ~4% of cells have an inter-active-origin distance >130 kb. Similar large inter-active-origin distances have been observed experimentally [45,46]. To prevent under-replication (or delay to cell cycle progression) these large inter-active-origin distances much be replicated prior to cell division. *S. cerevisiae* cells have about 60 min available to complete DNA replication before cell division starts, i.e. two forks converging at 2 kb/min can potentially replicate 240 kb during that time; therefore under-replication is unlikely according to our model.

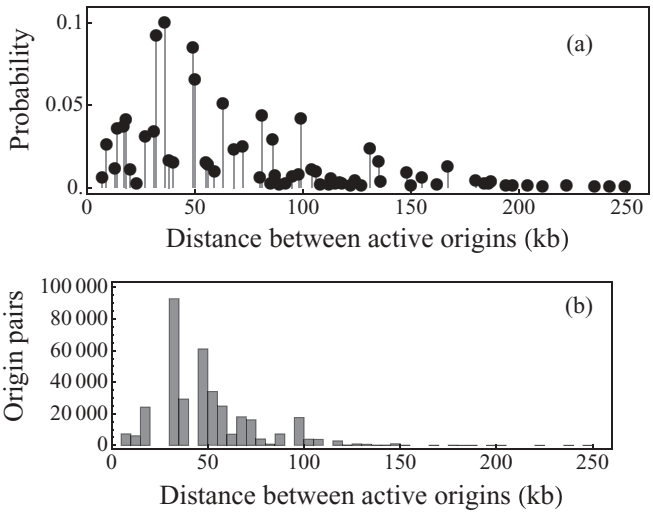

FIG. 4. Probability distribution of distances between active origins on *S. cerevisae* chromosome VI: (a) as a solution of Eq. (19), and (b) histogram for 10^5 Monte Carlo simulation. Parameter values from [4].

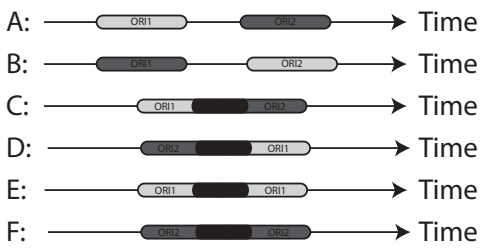

FIG. 5. A particular position along a chromosome in different cells can be replicated by forks from origin 1 (light gray), origin 2 (dark gray), or either origin (black). The order in which this happens (indicated in A–F above) as time progresses from left to right defines the possible states of replication.

IV. PARAMETER REGIMES

We want to use the general theory presented above to study replication dynamics in a simple setting. From now on we focus on the case of a hypothetical linear chromosome with just two origins. We define the chromosomal coordinates so that one of the origins has position $x_1 = 0$; the other origin has position $x_2 = D$. We assume that each origin can activate within a time window Δt_i with uniform probability; we will argue later that our conclusions are largely independent of the precise shape of the probability distribution. We select the time axis in such a way that the average activation time of the first origin is 0. The other origin has an average activation time τ, and we assume without loss of generality that $\tau \geqslant 0$. Thus the activation time distributions are

$$p_i(t) = \frac{q_i}{\Delta t_i} \quad \text{if} \quad t \in \left[t_i^{\text{av}} - \frac{\Delta t_i}{2}, t_i^{\text{av}} + \frac{\Delta t_i}{2} \right], \tag{20}$$

where $i = 1,2$, $t_1^{\text{av}} = 0$ and $t_2^{\text{av}} = \tau$, and p_1 and p_2 are set to zero outside the stated intervals.

Using Eqs. (6) and (20), we can write analytical expressions for the probability density $P(x,t)$ and other quantities of interest. From Eqs. (20) and (5), $P(x,t)$ vanishes outside the intervals $I_1 = [I_1^l, I_1^u]$ and $I_2 = [I_2^l, I_2^u]$ given by

$$I_i^l = \frac{|x_i - x|}{v} + t_i^{\text{av}} - \frac{\Delta t_i}{2};$$
$$I_i^u = \frac{|x_i - x|}{v} + t_i^{\text{av}} + \frac{\Delta t_i}{2}.$$

A. States of replication

There are six possible states of the replication at each position x on the chromosome, defined by the order in which forks from the two origins can replicate x. These are summarized in Fig. 5, where each position has a probability to be replicated at any time t by origin 1, origin 2, or by either origin depending on the distance from each origin and other parameters. The different possibilities are denoted by "states" A to F, as defined in Fig. 5.

State A corresponds to the situation where forks from origin 1 arrive at x before any fork coming from origin 2; if the competence q_1 of origin 1 is equal to 1, then this position is completely replicated by the first origin. For $q_1 < 1$, x is replicated by origin 2 only when origin 1 fails to activate. The replication probability density $P(x,t)$ is given in this case by

$$P(x,t) = \begin{cases} \frac{q_1}{\Delta t_1} & \text{if} \quad t \in I_1, \\ \frac{s_1 q_2}{\Delta t_2} & \text{if} \quad t \in I_2, \\ 0 & \text{otherwise.} \end{cases} \tag{21}$$

In similar manner, state B corresponds to forks from origin 2 arriving at x before forks from origin 1:

$$P(x,t) = \begin{cases} \frac{q_2}{\Delta t_2} & \text{if} \quad t \in I_2, \\ \frac{s_2 q_1}{\Delta t_1} & \text{if} \quad t \in I_1, \\ 0 & \text{otherwise.} \end{cases} \tag{22}$$

For the states C and D, there is an overlapping time window, when forks originating at either of the origins can get to x first. States E and F show that the time window where x is replicated by forks originating at one of the origins is contained in the time window of the other fork; this latter state is possible only if $\Delta t_1 \neq \Delta t_2$.

If we denote by $[t_l, t_u]$ the overlapping time window in cases C–F, then the replication probability densities for these four cases are given by the expression below, with i and j having different meanings for each state: C ($i = 1, j = 2$), D ($i = 2, j = 1$), E ($i = j = 1$), and F ($i = j = 2$):

$$P(x,t) = \begin{cases} p_i & \text{if} \quad t \in \left[I_i^l, t_l \right], \\ p_1 \left(\int_t^{I_2^u} p_2 dt + s_2 \right) + p_2 \left(\int_t^{I_1^u} p_1 dt + s_1 \right) & \\ \qquad \text{if} \quad t \in [t_l, t_u], & \\ p_j & \text{if} \quad t \in \left[t_u, I_j^u \right], \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

Expressions for t_l and t_u for the different states are given in the table below:

State	t_l	t_u
C	$\frac{\lvert D-x \rvert}{v} + \tau - \frac{\Delta t_2}{2}$	$\frac{\lvert x \rvert}{v} + \frac{\Delta t_1}{2}$
D	$\frac{\lvert x \rvert}{v} - \frac{\Delta t_1}{2}$	$\frac{\lvert D-x \rvert}{v} + \tau + \frac{\Delta t_2}{2}$
E	$\frac{\lvert D-x \rvert}{v} + \tau - \frac{\Delta t_2}{2}$	$\frac{\lvert D-x \rvert}{v} + \tau + \frac{\Delta t_2}{2}$
F	$\frac{\lvert x \rvert}{v} - \frac{\Delta t_1}{2}$	$\frac{\lvert x \rvert}{v} + \frac{\Delta t_1}{2}$

From the above table we can see that the state changes as position is changed along the chromosome. Based on these states it is possible to classify the replication dynamics into a number of parameter regimes. Figure 6 shows such regimes for $\Delta t_1 = \Delta t_2 = \Delta t$ and $\Delta t_1 \neq \Delta t_2$. In the first case [Fig. 6(a)], the number of parameters is reduced and we can look at the full dependencies between τ, Δt, D, and v. When $\Delta t_1 \neq \Delta t_2$, we will look at regimes for a few values of τ, Δt_1, and Δt_2 [Figs. 6(b)–6(d)].

B. Regimes for $\Delta t_1 = \Delta t_2$

In total there are five regimes (depending on the relative values of τ, Δt, D, and v) that differ in the dynamics of how

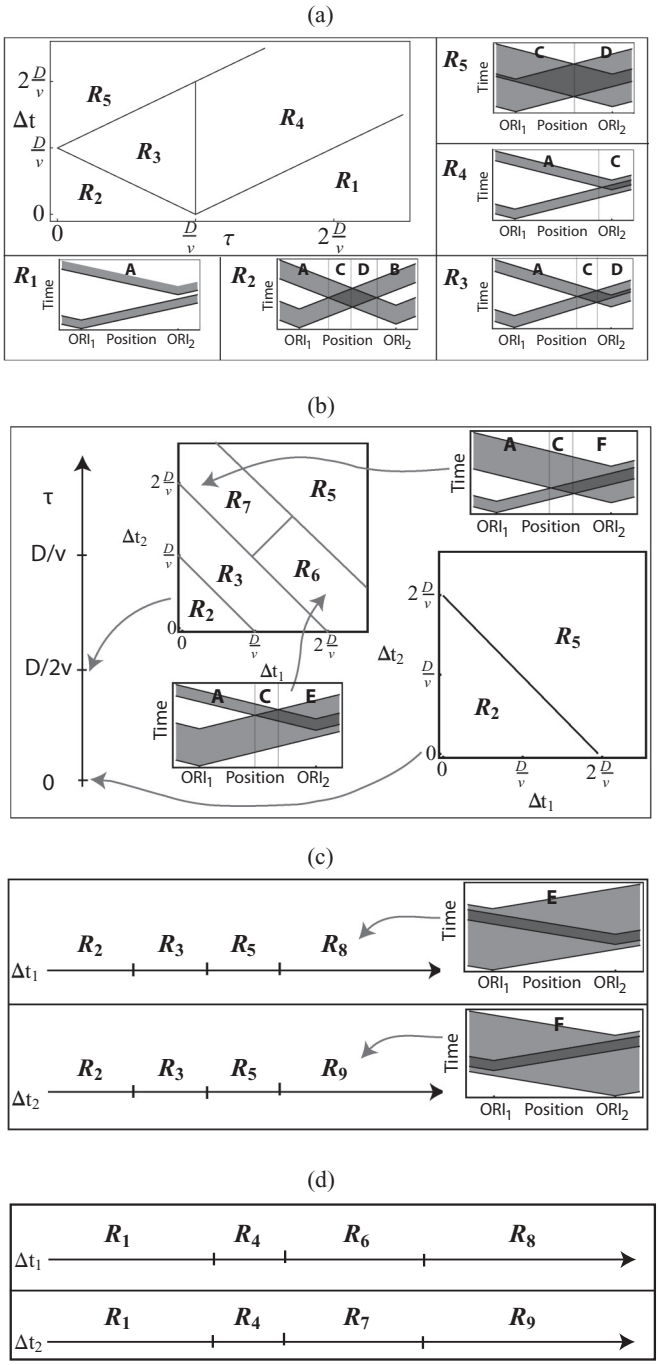

FIG. 6. Plots showing parameter regime diagrams based on the spatiotemporal dynamics of replication: gray areas indicate where the activation time distribution of origin 1 or 2 has nonzero values, and dark gray regions are where distributions for the two origins overlap. (a) Regimes for $\Delta t_1 = \Delta t_2$. (b) Some of the regimes for $\Delta t_1 \neq \Delta t_2$. (c) Regimes for variable Δt_1 (fixed value of $\Delta t_2 \ll D/v$) and Δt_2 (fixed value of $\Delta t_1 \ll D/v$) for $\tau < D/v$. (d) Regimes for variable Δt_1 (fixed value of $\Delta t_2 \ll D/v$) and Δt_2 (fixed value of $\Delta t_1 \ll D/v$) for $\tau > D/v$. Parameter values for (c) and (d) where replication dynamics undergoes changes are given in Appendix C.

the chromosome is replicated [shown in Fig. 6(a)]. For regime R_1 only state A occurs for the whole chromosome.

Most relevant to biological systems is regime R_2, since this is the case for many pairs of origins in real chromosomes. In this regime, the condition $\tau + \Delta t < D/v$ is satisfied, meaning that origins activate at a similar time and the variations in the activation time Δt are small enough that a fork from one origin can replicate the other origin only if that origin is not competent. Regime R_2 is formed from states A, B, C, and D. Changes occur at positions $x = \frac{1}{2}[D + v(\tau - \Delta t)]$ (from A to C), $x = \frac{1}{2}(D + v\tau)$ (from C to D), and $x = \frac{1}{2}[D + v(\tau + \Delta t)]$ (from D to B).

Regime R_2 describes the situation where two origins are sufficiently apart that they are not passively replicated by each other. It is quite common, however, to find two origins located close to each other in a chromosome, such that the condition $\tau + \Delta t < D/v$ is violated. In this case, regimes other than R_2 describe the replication dynamics of the system.

In regime R_3, in some cells origin 2 actives just before forks from origin 1 arrive at position $x = x_2 = D$ and in some cells both origins compete to replicate positions on the right of $x = \frac{1}{2}[D + v(\tau - \Delta t)]$. At this point the behavior changes from state A to C.

For regime R_4 state A occurs up to position $x = \frac{1}{2}[D + v(\tau - \Delta t)]$, then it changes to state C. In a similar way, regime R_5 changes from state C to state D at position $x = \frac{1}{2}(D + v\tau)$. More details on the dynamics of replication for different sets of parameters are shown in Fig. 6(a).

For a fixed value of $\Delta t < D/v$, as the difference in average activation time between origins, τ, increases, the replication dynamics undergoes the following changes:

Change	τ	Change	τ	Change	τ
$R_2 \to R_3$	$\frac{D}{v} - \Delta t$	$R_3 \to R_4$	$\frac{D}{v}$	$R_4 \to R_1$	$\frac{D}{v} + \Delta t$

For any fixed $\tau < D/v$, increasing the width Δt of the activation time distribution leads to change of the regime from R_2 to R_3 at $\frac{D}{v} - \tau$ and from R_3 to R_5 at $\frac{D}{v} + \tau$.

A discussion on regimes for $\Delta t_1 \neq \Delta t_2$ is given in Appendix C.

V. DYNAMICS OF REPLICATION FORKS

The proportion of left and right moving forks will depend on the properties of the origins and fork velocity, as indicated by Eq. (17). The analytical expression for the proportion of left moving forks, n_{left}, for regime R_2, is given by Eq. (C1) in Appendix C. In regime R_2 the proportion of left moving forks takes a sigmoidal shape. Figures 7(a), 7(c), and 7(e) show the effect of varying parameters on proportion of left moving forks in regime R_2: as a function of competence q_2 [Fig. 7(a)], as a function of time τ [Fig. 7(c)], and as a function of the width of the activation time distribution $\Delta t_1 = \Delta t_2 = \Delta t$ [Fig. 7(e)]. Parameter values are $q_1 = q_2 = 1$, $\tau = 0$, $\Delta t_1 = 5$, $\Delta_2 = 5$, and $v = 1$, if not otherwise stated. The consequences of differences in parameter values can be clearly distinguished: with decreasing competence, the curve is pushed down; increasing τ shifts the curve closer to the second origin; decreasing Δt increases the gradient of the sigmoidal function.

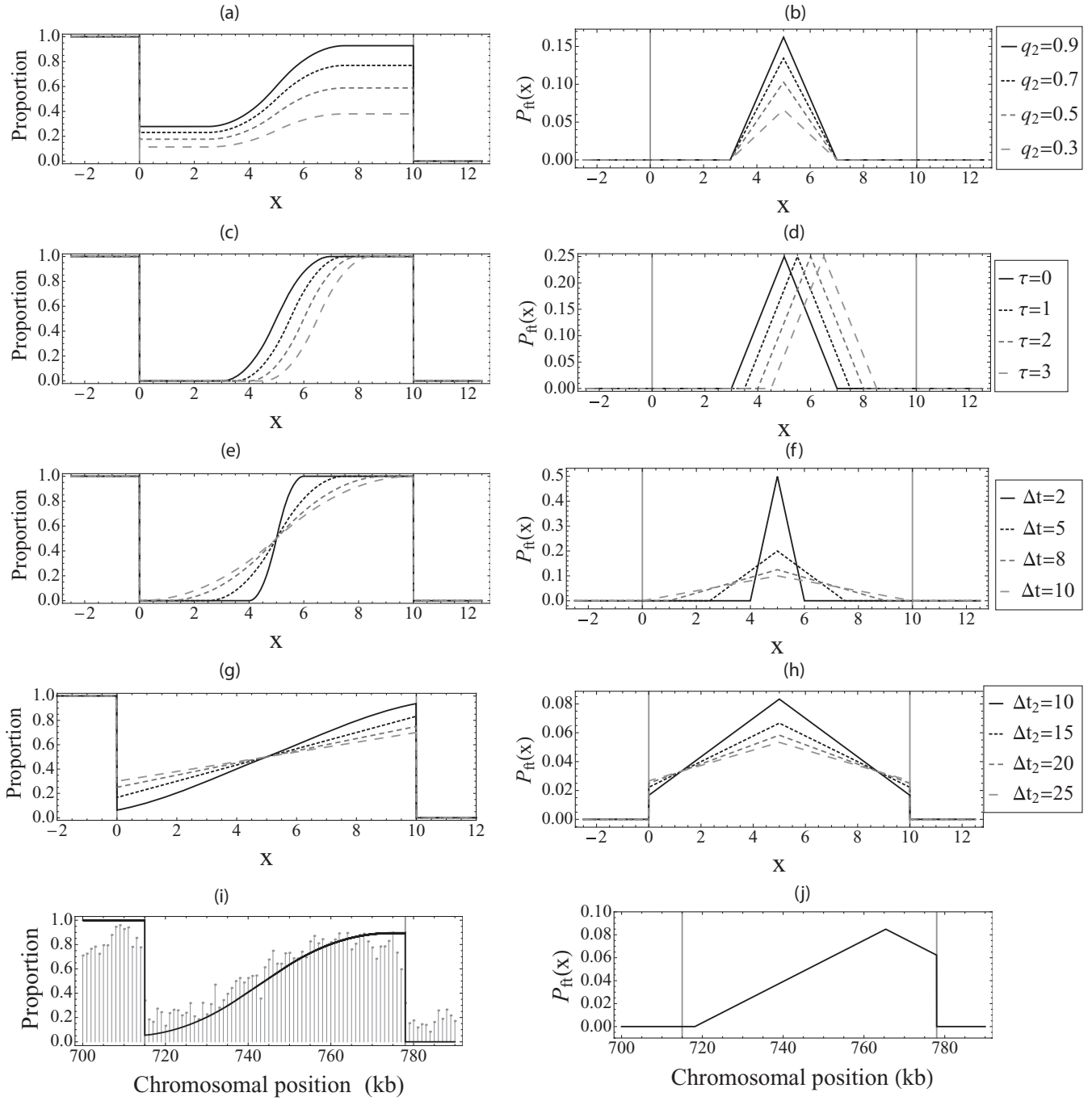

FIG. 7. The effect of varying parameters on the proportion of left moving forks and fork termination: (a),(b) as a function of competence q_2, (c),(d) as a function of time τ, (e),(f) as a function of width of activation $\Delta t_1 = \Delta t_2 = \Delta t$, and (g),(h) as a function of width of activation Δt_2. (i) Experimentally measured proportion of left moving forks (gray) [47] on the *S. cerevisiae* chromosome VII (700–780 kb) with a curve based on Eq. (17) (black) and (j) prediction of the fork termination position probability distribution based on Eq. (D1).

We have also looked at other regimes—in Fig. 7(g) the parameter values $\Delta t_2 = 10$ and $\Delta t_2 = 15$ correspond to R_7, while $\Delta t_2 = 20$ and $\Delta t_2 = 25$ correspond to R_9; here $q_1 = q_2 = 1$, $\tau = 0$ and $\Delta t_1 = 5$. Unlike regime R_2, these regimes permit passive replication of one (R_7) or both (R_9) origins. It can be seen that for these regimes the proportion of left moving forks between the two origins becomes linear with respect to position on the chromosome.

Recently, high-resolution analysis of Okazaki fragment synthesis was performed in *S. cerevisiae* [47]; this allows determination of the lagging strand proportion and hence the proportion of left moving forks. Figure 7(i) shows the

experimentally determined proportion of left moving forks for a part of chromosome VII from 700 to 780 kb (gray lines) with a curve based on Eq. (17) (black) [48]. This demonstrates the close agreement between our model prediction and an independent experimental data set.

VI. RESULTS FOR THE FORK TERMINATION POSITION PROBABILITY DISTRIBUTION

Although the forks all start at the same locations (the origins), they meet each other and terminate at different locations in each cell, because of the stochastic activation times. Only forks traveling in opposite directions between the two origins can collide. This condition limits positions of fork termination to the interval $[0; D]$ (excluding the two ends of the chromosome). (If one of origins is activated much later in time, i.e., the dynamics is in regime R_1, only one origin activates and therefore there are no termination events between the two origins.) The fork termination position distribution will have nonzero values in the areas of the space-time diagram where the probability of that position being replicated by forks from both origins is nonzero. An analytical expression for the fork termination position probability distribution function (PDF) is given in Appendix D.

Figures 7(b), 7(d), 7(f), and 7(h) illustrate the effect of varying the parameters q_2, τ, Δt, and Δt_2. For $\tau = 0$, the maximum of the PDF is at position $x = \frac{D}{2}$ and takes the value $p^* = \frac{q_1 q_2 (\Delta t_1 + \Delta t_2)}{2\Delta t_1 \Delta t_2 (1 - s_1 s_2)}$. If τ is increased, the maximum moves to the right by $\frac{v\tau}{2}$. The area under the PDF is $\frac{(\Delta t_1 + \Delta t_2)^2 q_1 q_2 v}{8\Delta t_1 \Delta t_2 (1 - s_1 s_2)}$ for values $\tau < \frac{D}{v} - \frac{\Delta t_1 + \Delta t_2}{2}$; for increasing τ it starts decreasing until zero is reached at $\tau = \frac{2D + v(\Delta t_1 + \Delta t_2)}{2v}$.

The biological literature commonly suggests that termination sites are disperse, but there is limited evidence for discrete termination sites [39]. Our model indicates that termination sites have a continuous distribution. Equation (D1) can help identify the most probable positions of fork termination sites on the chromosome. Figure 7(j) shows the prediction of the fork termination position probability distribution for a part of chromosome VII from 700 to 780 kb.

VII. SIGNATURES OF STOCHASTICITY IN REPLICATION PROFILES

In this section we will discuss how information about the stochasticity of the activation times of origins—the width of the activation window—can be obtained from the mean replication time as a function of chromosomal position, $T(x)$. Because the parameter space in the generic case is very big, from now on we will consider only the case $\Delta t_1 = \Delta t_2 = \Delta t$.

An analytical expression for $T(x)$ for the case where the condition $\tau + \Delta t < D/v$ is satisfied has been derived [5]. This corresponds to regime R_2 and is the case for many pairs of adjacent origins in real chromosomes. This is valid when the variations in the activation time Δt are small enough that a fork from one origin can replicate the other origin only if that origin is not competent.

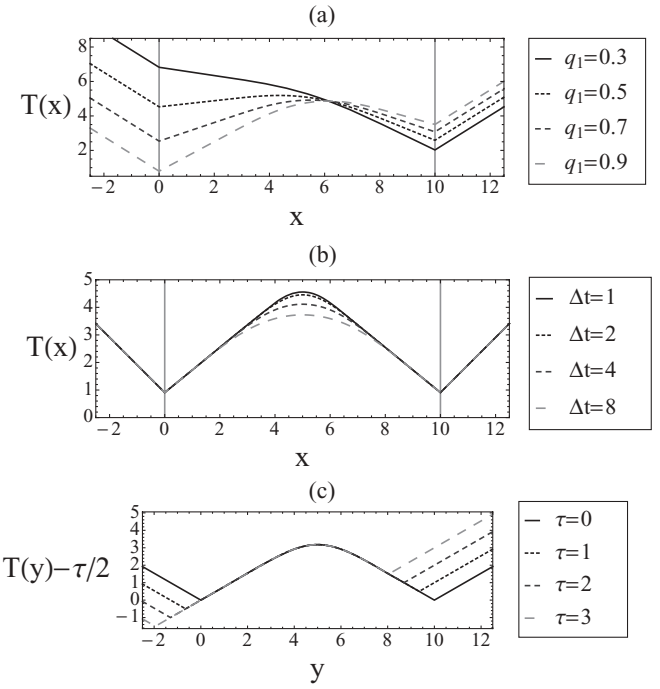

FIG. 8. Replication time curves for (a) differing values of competence q_1; (b) different widths of the activation time window; (c) different values of τ, with position $y = x + v\tau/2$.

A plot of $T(x)$ for various sets of parameters is shown in Fig. 8. We see that $T(x)$ has discontinuous derivatives at the origin locations. This is due to forks originating only at the origin locations; there is a discontinuous change in the proportion of left propagating compared to right propagating forks as one crosses an origin site. At the origins, the mean replication times are [5]

$$\begin{aligned} T(x_1) &= T(0) = q_2 s_1 (D/v + \tau)/(1 - s_1 s_2); \\ T(x_2) &= T(D) = (q_1 s_2 D/v + q_2 \tau)/(1 - s_1 s_2). \end{aligned} \tag{24}$$

It is commonly assumed in the replication literature that $T(x)$ has a minimum at an origin, and that the value of this minimum directly gives the average activation time for the origin. However, Eq. (24) shows that this is not the case and, in fact, $T(x_i) \geqslant t_i^{\mathrm{av}}$: the mean replication time at an origin location is equal to or greater than the origin's average activation time. Only when an origin has $q_i = 1$ can $T(x_i) = t_i^{\mathrm{av}}$, because if an origin fails to activate in a given cell, the DNA at the origin location will not be replicated until a fork from another origin arrives. This means that T_i is higher for origins that are more likely to fail, as seen directly in Fig. 8(a). Another important conclusion is that even when both origins have the same average activation time ($\tau = 0$), generally we have $T(x_1) \neq T(x_2)$, also shown in [5]. This is again due to the possibility of origins not activating. Therefore, the origin with the lower minimum of $T(x)$ does not necessarily activate earlier than the other origin: minima of $T(x)$ cannot be used to draw conclusions on the relative activation times of the corresponding origins, as previously assumed [9,16]. The equation for mean replication time shows that in general $T(x)$ at any point depends collectively on the parameters of

all origins. However, if an origin is highly competent, early activating, and isolated from other origins, $T(x)$ at that origin's position will be close to the origin's average activation time.

The expression for mean replication time, Eq. (24), challenges the assumption that there is a one-to-one correspondence between replication origins and minima of $T(x)$: minima correspond to origin locations, but there may be origins for which there is no peak. The slope of T near the first origin (for $x > 0$) is [5]

$$T'(x) = \frac{q_1 - q_2 s_1}{v(1 - s_1 s_2)}. \tag{25}$$

This expression shows that the slope is a function of the competences q_i of both origins as well as the fork velocity v. For the origin at $x = 0$ to be a minimum of $T(x)$, we must have $T' > 0$ for $x > 0$, from which we get the condition

$$q_1 > \frac{q_2}{1 + q_2}. \tag{26}$$

In a similar way, the slope of $T(x)$ near the second origin (for $x < D$) is

$$T'(x) = \frac{q_1 s_2 - q_2}{v(1 - s_1 s_2)}, \tag{27}$$

which yields the condition

$$q_2 > \frac{q_1}{1 + q_1}. \tag{28}$$

This shows that if an origin has low competence compared to its neighbor, it may not be a minimum of $T(x)$, as illustrated in Fig. 8(a). This phenomenon has been observed in experimental data [16]. Note that if $q_1 > 1/2$, this condition is always satisfied and a minimum is guaranteed for this two-origin system and this regime. Figure 9 shows values for q_1 and q_2, for which Eqs. (26) and (28) are satisfied, i.e., both origins are local minima in the replication time curve. In addition, Eqs. (25) and (27) show that the fork velocity is not given by the slope of $T(x)$, an assumption widely used in the literature [16].

In contrast to the discontinuity of $T'(x)$ at the origin locations, the plots of replication time curves show that the local maximum of $T(x)$ between two origins is a smooth curve. The reason is that in different cells in a population forks meet each other and terminate at different locations on the DNA, as has been shown in Sec. VI. This suggests that the shape of the maximum of $T(x)$, when it exists, could be used to infer information about the width of the activation window. We expect that sharp maxima should correspond to forks meeting within a narrow time window; conversely, a broad maximum corresponds to a high time window. This can be seen in Fig. 8(b), where $T(x)$ is plotted for various values of Δt.

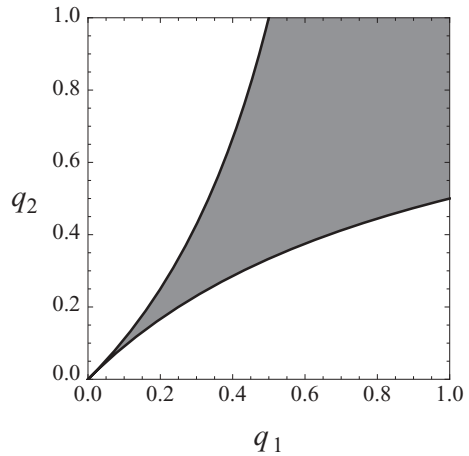

FIG. 9. Gray area indicates the sets of competencies q_1 and q_2 for which both origins are local minima in the replication time curve.

In order to investigate this more quantitatively, we use the modulus of the second derivative of $T(x)$ at the maximum to measure how broad the maximum is, because low values of $|T''(x)|$ correspond to broad peaks. At the maximum of $T(x)$ [5],

$$|T''(x^*)| = \frac{4q_1 q_2 \sqrt{1 - \left|\frac{1}{q_1} - \frac{1}{q_2}\right|}}{v^2 \Delta t (1 - s_1 s_2)}, \tag{29}$$

where

$$x^* = \begin{cases} \frac{D + v\tau}{2} & \text{for } q_1 = q_2; \\ \frac{D + v\tau}{2} + \frac{v\Delta t}{2}\left(1 - \left|\frac{1}{q_1} - \frac{1}{q_2}\right|\right) & \text{for } q_1 > q_2; \\ \frac{D + v\tau}{2} - \frac{v\Delta t}{2}\left(1 - \left|\frac{1}{q_1} - \frac{1}{q_2}\right|\right) & \text{for } q_1 < q_2. \end{cases}$$

Thus $|T''(x^*)|$ is inversely proportional to Δt. Notice also that $|T''(x^*)|$ does not depend on τ, which means it is independent of the origins' average activation times. Figure 8(c) shows replication time curves for different values of τ, where the position along the chromosome has been shifted by $y = x + \frac{v\tau}{2}$ and replication times are shifted downwards by $\frac{\tau}{2}$. This figure clearly shows that the shape of $T(x)$ near maximum is independent of τ.

Expression (29) can be used to calculate Δt from an experimental replication time profile $T(x)$, if the origin competences and the fork velocity are known. This is a very useful result because it allows the determination of a quantity characterizing stochastic properties of the system, Δt, from $T(x)$, which is defined by a population average. This is valuable because experiments to directly measure Δt are technically difficult [17,18]. We note that this does not require assuming that all cells in the population are synchronized, since in each individual cell in an asynchronous population, the statistics of the relative activation times of origins remains unaltered [4].

VIII. GAUSSIAN AND SKEWED ORIGIN ACTIVATION DISTRIBUTIONS

Although Eq. (29) was obtained using a simple uniform distribution for $p_i(t)$, we expect it to be a good approximation for any single-peaked distribution function $p_i(t)$, since Eq. (29) involves only the second moment (the variance) of the distribution, and the replication dynamics are mostly determined by the average activation time and the width of the activation distribution. So we expect two single-peaked distributions with the same t_{av} and Δt to have very similar $T''(x)$.

To test this assumption, we used Eq. (6) to numerically compute $T(x)$ for pairs of origins with Gaussian activation

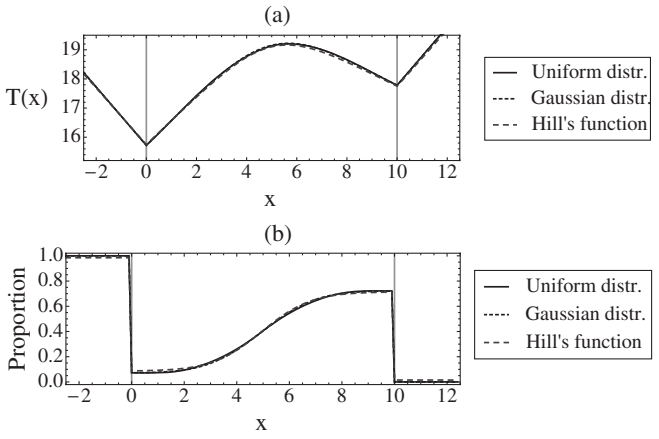

FIG. 10. Plots showing mean replication time (a) and proportion of left moving forks (b) for three different distributions: Uniform distribution ($t_i^{\rm av} = 15, \Delta t_i = 8$), Gaussian distribution ($\mu_i = 15, \sigma_i = 2.31$), and skewed distribution given by Eq. (30) ($t_{1/2i} = 15.35, t_{wi} = 2.77$), $i = 1,2$. For all three cases $q_1 = 0.9, q_2 = 0.7, v = 1$.

time distribution and with a skewed distribution leading to a sigmoidal cumulative activation time distribution, described by a Hill's-type function

$$F(t) = \frac{q \ln 3 t_{1/2} \left(\frac{t_{1/2}}{t}\right)^{\ln 3/\kappa - 1}}{\kappa t^2 \left[1 + \left(\frac{t_{1/2}}{t}\right)^{\ln 3/\kappa}\right]^2}, \tag{30}$$

where

$$\kappa = \ln\left(\frac{t_w + \sqrt{4t_{1/2}^2 + t_w^2}}{2t_{1/2}}\right), \tag{31}$$

as used in Ref. [14].

Figure 10 shows good agreement of the mean replication time and proportion of left moving forks for the three distributions mentioned above. Choosing parameters for which all these distributions have the same mean and variance, we find that in all cases $T''(x)$ never differs between distributions by more than 10%. This means that Eq. (29) is a very accurate prediction of Δt, regardless of the detailed shape of $p_i(t)$.

IX. APPLICATION TO EXPERIMENTAL DATA

We expect Eq. (29) to be a reliable prediction for isolated pairs of origins whose competences are not too low, so that there are well-defined peaks at the origin positions; this ensures that most forks traveling in the region between the two origins come from those origins, which is what is required for Eq. (29) to hold. We also note that there are organisms with far fewer replication origins than *S. cerevisiae*, for which the two-origin assumption involves little or no approximation; in particular, many archaea have only two or three origins, and high-throughput methods to study their replication dynamics are available [49–51].

We will apply the above model to data from *S. cerevisiae* (brewer's yeast), which has 16 chromosomes and about 300 origins [52]. A rough criterion for a pair of origins to be considered "isolated" is if the distance between them is smaller

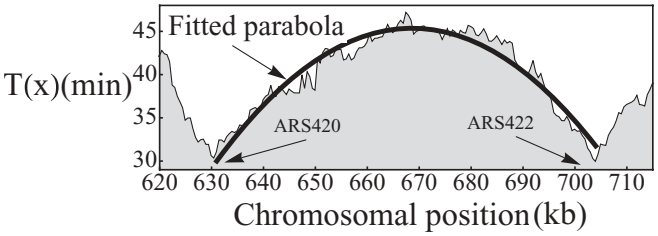

FIG. 11. Replication time curves [16] for *S. cerevisiae* chromosome IV (620–715 kb) with a fitted parabola

than the distance between either origin and its immediate neighbor. More precisely, let (x_i, x_{i+1}) be a pair of origins in a chromosome, and $d_{i,i+1}$ be the distance between the origin at x_i and the origin at x_{i+1}. Then our criterion for the pair (x_i, x_{i+1}) to be isolated is that it satisfies both $d_{i,i+1} < d_{i-1,i}$ and $d_{i,i+1} < d_{i+1,i+2}$. We have used the DNA replication origin database [52] to compute how many pairs of *S. cerevisiae* organism satisfy this criterion, and we found that 35% do. Using the stochastic simulation procedure described in Ref. [14], we found that of the origin pairs which are isolated, in 40% of those the Δt's of the two origins are within 2 min of each other (meaning 14% of the origins overall). If we relax the definition of "close" to a 3 min difference, this fraction increases to 60% (or 21% overall).

We looked at experimental data [16] for *S. cerevisiae* chromosome IV (the region containing origins $ARS420$ and $ARS422$), shown in Fig. 11. The smoothness of the curve—ignoring the fluctuations caused by experimental noise—is direct evidence for stochastic origin activation, in agreement with other results [17,18]. We fitted a parabola through the data points and from this determined the value of $|T''|$. Using Eq. (29) we estimate the values of Δt as 15 min [53]. This value is in agreement with the limited number of single cell measurements that have been made at other *S. cerevisiae* origins [18].

X. DISCUSSION AND CONCLUSIONS

In this paper we studied the properties of a mathematical model of chromosome replication that describes genome duplication dynamics in eukaryotic cells. In the first part of the study, we formulated a general model for replication, which takes into account the fact that origins activate stochastically, and also accounts for the possibility that origins in individual cells may not activate at all, because they may have failed to license. We have derived a probability distribution of replication as a function of position and time, and from this we have found analytical expressions for many quantities of interest such as mean replication time, the average number of active origins, and others.

In the second part of the paper, the dynamics of replication was investigated in detail in an idealized scenario for a chromosome with two origins of replication and the activation probability distribution given by a uniform distribution. Despite being idealized, this simple scenario encapsulates much of the essential behavior found in more complex cases; and although organisms vary enormously in the size, shape, and

distribution of genomes, the case of the chromosome with just two origins is of biological interest: recently an *E. coli* mutant with two identical functional replication origins has been constructed [54]. We were able to derive an analytical expression for the mean replication time and to extract hence many important properties of the dynamics of DNA replication. Furthermore, we have proposed a method to determine the width of the activation time probability distribution from experimentally measured population-averaged data. Our results compare favorably with experimental measurements in *S. cerevisae*.

ACKNOWLEDGMENTS

We thank M. Hawkins for valuable discussions. This work has been supported through the Biotechnology and Biological Sciences Research Council (Grants No. BB/E023754/1, No. BB/G001596/1, and No. BB-G010722).

APPENDIX A: RELATIONSHIP BETWEEN MEAN REPLICATION TIME AND MEAN COPY NUMBER

For a eukaryotic cell with a fixed S phase, the mean replication time can be calculated:

$$T(x) = \frac{1}{\mathcal{N}} \int_{T_1}^{T_2} t\, P(x,t) dt, \tag{A1}$$

where $\mathcal{N} = 1 - \prod_{i=1}^{n} s_i$, T_1 is the time when replication starts, and T_2 is the time when replication finishes.

The relationship between copy number $C(x,t)$ and fraction replicated $m(x,t)$ is given by Eq. (8). Now, the fraction replicated before time t for cells where replication started at T_1 is given by

$$m(x,t) = \frac{1}{\mathcal{N}} \int_{T_1}^{t} P(x,\tau) d\tau. \tag{A2}$$

The time average of the function $m(x,t)$ is

$$\begin{aligned}
\bar{m}(x) &= \frac{1}{\mathcal{N}} \frac{1}{T_2 - T_1} \int_{T_1}^{T_2} dt \int_{T_1}^{t} P(x,\tau) d\tau \\
&= \frac{1}{\mathcal{N}} \frac{1}{T_2 - T_1} \int_{T_1}^{T_2} d\tau \int_{\tau}^{T_2} P(x,\tau) dt \\
&= \frac{1}{\mathcal{N}} \frac{1}{T_2 - T_1} \int_{T_1}^{T_2} P(x,\tau) d\tau \int_{\tau}^{T_2} dt \\
&= \frac{1}{\mathcal{N}} \frac{1}{T_2 - T_1} \int_{T_1}^{T_2} P(x,\tau)(T_2 - \tau) d\tau \\
&= \frac{1}{\mathcal{N}} \frac{1}{T_2 - T_1} \left(\int_{T_1}^{T_2} P(x,\tau) T_2 d\tau - \int_{T_1}^{T_2} \tau\, P(x,\tau) d\tau \right) \\
&= \frac{1}{\mathcal{N}} \frac{1}{T_2 - T_1} \left(T_2 \int_{T_1}^{T_2} P(x,\tau) d\tau - \int_{T_1}^{T_2} t\, P(x,t) dt \right).
\end{aligned}$$

Taking into account Eq. (A1) and the fact that the first integral in the parentheses equals $1 - \prod_{i=1}^{n} s_i = \mathcal{N}$, we get

$$\bar{m}(x) = \frac{T_2 - T(x)}{T_2 - T_1} \tag{A3}$$

and

$$\bar{C}(x) = \frac{T_2 - T(x)}{T_2 - T_1} + 1. \tag{A4}$$

APPENDIX B: REGIMES FOR $\Delta t_1 \neq \Delta t_2$

Some of the regimes for the case $\Delta t_1 \neq \Delta t_2$ are shown in Fig. 6(b). Only in this case can the states E and F (Fig. 5) be observed. Also, regimes R_2, R_3, and R_5 have either state E occurring between positions $x = \frac{1}{4}[2D + v(2\tau - \Delta t_1 + \Delta t_2)]$ and $x = \frac{1}{4}[2D + v(2\tau + \Delta t_1 - \Delta t_2)]$ for $\Delta t_1 > \Delta t_2$, or state F between positions $x = \frac{1}{4}[2D + v(2\tau + \Delta t_1 - \Delta t_2)]$ and $x = \frac{1}{4}[2D + v(2\tau - \Delta t_1 + \Delta t_2)]$ for $\Delta t_1 < \Delta t_2$. If $\Delta t_1 \to \Delta t_2$, the length of this interval decreases until states E and F disappear and position $x = \frac{1}{2}(D + v\tau)$ then separates states C and D.

For regime R_6, state A occurs up to position $x = \frac{1}{4}[2D + v(2\tau - \Delta t_1 - \Delta t_2)]$; then it changes to state C. At position $x = \frac{1}{4}[2D + v(2\tau - \Delta t_1 + \Delta t_2)]$, state C changes to the state E. In a similar way, for regime R_7 state A also occurs up to position $x = \frac{1}{4}[2D + v(2\tau - \Delta t_1 - \Delta t_2)]$, and then it changes to state C. At position $x = \frac{1}{4}[2D + v(2\tau + \Delta t_1 - \Delta t_2)]$, state C changes to the state F.

Other types of regime characteristic for the case $\Delta t_1 \neq \Delta t_2$ are regime R_8 ($\Delta t_1 \gg \Delta t_2$), where only state E is possible; and regime R_9 ($\Delta t_2 \gg \Delta t_1$), where only state F occurs.

For a fixed value of $\Delta t_1 \ll \frac{D}{v}$ the regime dynamics as Δt_2 changes are shown in the upper panels of Figs. 6(c) and 6(d). The transitions in the dynamics are given in the table below:

	$\tau < \frac{D}{v}$		$\tau > \frac{D}{v}$
Change	Δt_1	Change	Δt_1
$R_2 \to R_3$	$2\left(\frac{D}{v} - \tau - \frac{\Delta t_2}{2}\right)$	$R_1 \to R_4$	$2\left(\tau - \frac{D}{v} - \frac{\Delta t_2}{2}\right)$
$R_3 \to R_5$	$2\left(\frac{D}{v} - \tau + \frac{\Delta t_2}{2}\right)$	$R_4 \to R_6$	$2\left(\tau - \frac{D}{v} + \frac{\Delta t_2}{2}\right)$
$R_5 \to R_8$	$2\left(\tau + \frac{D}{v} + \frac{\Delta t_2}{2}\right)$	$R_6 \to R_8$	$2\left(\frac{D}{v} + \tau + \frac{\Delta t_2}{2}\right)$

For a fixed value of $\Delta t_2 \ll \frac{D}{v}$ the regime dynamics for variable Δt_1 is shown in the lower panels of Figs. 6(c) and 6(d). The transitions are

	$\tau < \frac{D}{v}$		$\tau > \frac{D}{v}$
Change	Δt_2	Change	Δt_2
$R_2 \to R_3$	$2\left(\frac{D}{v} - \tau - \frac{\Delta t_1}{2}\right)$	$R_1 \to R_4$	$2\left(\tau - \frac{D}{v} - \frac{\Delta t_1}{2}\right)$
$R_3 \to R_5$	$2\left(\frac{D}{v} - \tau + \frac{\Delta t_1}{2}\right)$	$R_4 \to R_7$	$2\left(\tau - \frac{D}{v} + \frac{\Delta t_1}{2}\right)$
$R_5 \to R_9$	$2\left(\frac{D}{v} + \tau + \frac{\Delta t_1}{2}\right)$	$R_7 \to R_9$	$2\left(\frac{D}{v} + \tau + \frac{\Delta t_1}{2}\right)$

In the extreme case where $\tau > \frac{D}{v} + \frac{\Delta t_1}{2} + \frac{\Delta t_2}{2}$ the only possible regime is R_1.

APPENDIX C: PROPORTION OF LEFT MOVING FORKS IN REGIME R_2

This is given by

$$n_{\text{left}}(x) = \begin{cases} 1 \quad \text{if} \quad x < 0; \\ \frac{q_2 - q_1 q_2}{q_1 + q_2 - q_1 q_2} \quad \text{if} \quad x \in \left[0; \frac{D+v(\tau-\Delta t)}{2}\right); \\ -\frac{q_2\{D^2 q_1 - \Delta t^2(-2+q_1)v^2 - 2Dq_1(\Delta t v + 2x - v\tau) - 2\Delta t q_1 v(-2x+v\tau) + q_1(-2x+v\tau)^2\}}{2\Delta t^2\{q_1(q_2-1)-q_2\}v^2} \quad \text{if} \quad x \in \left[\frac{D+v(\tau-\Delta t)}{2}; \frac{D+v\tau}{2}\right); \\ \frac{q_2\{D^2 q_1 + \Delta t^2(-2+q_1)v^2 + 2\Delta t q_1 v(-2x+v\tau) + q_1(-2x+v\tau)^2 + 2Dq_1(-2x+v\tau+v\Delta t)\}}{2\Delta t^2\{q_1(q_2-1)-q_2\}v^2} \quad \text{if} \quad x \in \left[\frac{D+v\tau}{2}; \frac{D+v(\tau+\Delta t)}{2}\right); \\ \frac{q_2}{q_1+q_2-q_1 q_2} \quad \text{if } x \in \left[\frac{D+v(\tau+\Delta t)}{2}; D\right); \\ 0 \quad \text{if} \quad x \geqslant D. \end{cases} \tag{C1}$$

APPENDIX D: PROBABILITY OF THE FORK TERMINATION POSITION DISTRIBUTION

This is given by

$$P_{\text{ft}}(x) = \begin{cases} 0 \quad \text{if } \tau > \frac{2D+v(\Delta t_1+\Delta t_2)}{2v}; \\ 0 \quad \text{if } x < 0; \\ 0 \quad \text{if } x > D; \\ \frac{q_1 q_2[4x-2D+v(-2\tau+\Delta t_1+\Delta t_2)]}{2\Delta t_1 \Delta t_2 v(1-s_1 s_2)} \quad \text{if} \quad x \in \left[\frac{2D+v(2\tau-\Delta t_1-\Delta t_2)}{4}; \frac{D+v\tau}{2}\right]; \\ \frac{q_1 q_2[4x+2D+v(-2\tau+\Delta t_1+\Delta t_2)]}{2\Delta t_1 \Delta t_2 v(1-s_1 s_2)} \quad \text{if} \quad x \in \left[\frac{D+v\tau}{2}; \frac{2D+v(2\tau+\Delta t_1+\Delta t_2)}{4}\right]. \\ 0 \quad \text{otherwise.} \end{cases} \tag{D1}$$

APPENDIX E: SENSITIVITY ANALYSIS OF AN EXPRESSION FOR THE SECOND DERIVATIVE OF THE MEAN REPLICATION TIME

We rewrite Eq. (29) as $G = \frac{A}{g}$, where $A = \frac{4q_1 q_2 \sqrt{1-|1/q_1 - 1/q_2|}}{v^2(1-s_1 s_2)}$, and $g = \Delta t$.Then the change in the value of G with respect to differences in the parameter g is

$$|\Delta G| = A\left(\frac{1}{g} - \frac{1}{g+\Delta g}\right) = A\frac{\Delta g}{g(g+\Delta g)}$$

For the parameter g in the range (7,18) min [14], a difference of 2 min results in a difference of up to 2.5% in $\frac{\Delta G}{A}$.

Normalization errors and noisy data result in errors in the calculated mean replication time $[T(x)]$. We have used a simple Monte Carlo simulation technique to roughly estimate how the Δt predicted from formula (29) is affected by errors in the replication time profile $T(x)$. For this simulation we added Gaussian-distributed noise of mean 0 and standard deviation 5 min to each simulated replication profile of a two-origin chromosome, thus mimicking the effects of experimental error (we expect 5 min to be a reasonable estimate of the error for the experiments we use). We then numerically calculate the second derivative of a fitted parabola through the resulting mean replication time curve, and use formula (29) to calculate the "predicted" Δt^*, which we can compare with the actual value Δt used to generate $T(x)$ in the first place. We then repeat this procedure 100 times, generating a distribution of Δt^*'s, and we do this for a few values of Δt to see how the error affects origins with different variabilities in activation time.

Errors in the estimation of Δt are greatest for origins with low variation in the activation time, as is to be expected. For origins with $\Delta t \approx 10$ min, which is the value we estimated in the example discussed in the paper, the error is around 10%, which shows that formula (29) is quite robust to experimental errors.

[1] J. Smith and L. Martin, Proc. Natl. Acad. Sci. USA **70**, 1263 (1973).

[2] J. J. Blow and P. J. Gillespie, Nat. Rev. Cancer **8**, 799 (2008).

[3] Z. Smith and D. Higgs, Hum. Mol. Genet. **8**, 1373 (1999).

[4] A. P. S. de Moura, R. Retkute, M. Hawkins, and C. A. Nieduszynski, Nucleic Acids Res. **38**, 5623 (2010).

[5] R. Retkute, C. A. Nieduszynski, and A. de Moura, Phys. Rev. Lett. **107**, 068103 (2011).

[6] O. Hyrien and A. Goldar, Chromosome Res. **18**, 147 (2010).

[7] S. Jun, H. Zhang, and J. Bechhoefer, Phys. Rev. E **71**, 011908 (2005).

[8] J. Lygeros, K. Koutroumpas, S. Dimopoulos, I. Legouras, P. Kouretas, C. Heichinger, P. Nurse, and Z. Lygerou, Proc. Natl. Acad. Sci. USA **106**, 9535 (2008).

[9] T. Spiesser, E. Klipp, and M. Barberis, Mol. Genet. Genom. **282**, 25 (2009).

[10] M. D. Sekedat, D. Fenyö, R. S. Rogers, A. J. Tackett, J. D. Aitchison, and B. T. Chait, Mol. Syst. Biol. **6**, 353 (2010).
[11] O. Hyrien and A. Goldar, Chromosome Res. **18**, 147 (2009).
[12] K. Koutroumpas and J. Lygeros, Automatica **47**, 1156 (2011).
[13] H. Luo, J. Li, M. Eshaghi, J. Liu, and R. K. M. Karuturi, BMC Bioinf. **11**, 297 (2010).
[14] S. C.-H. Yang, N. Rhind, and J. Bechhoefer, Mol. Syst. Biol. **6**, 404 (2010).
[15] A. Baker, B. Audit, C.-L. Chen, B. Moindrot, A. Leleu, G. Guilbaud, A. Rappailles, C. Vaillant, A. Goldar, F. Mongelard, Y. d'Aubenton Carafa, O. Hyrien, C. Thermes, and A. Arneodo, PLoS Comput. Biol. **8**, e1002443 (2012).
[16] M. K. Raghuraman, E. A. Winzeler, D. Collingwood, S. Hunt, L. Wodicka, A. Conway, D. J. Lockhart, R. W. Davis, B. J. Brewer, and W. L. Fangman, Science **294**, 115 (2001).
[17] S. Tuduri, H. Tourriere, and P. Pasero, Chromosome Res. **18**, 91 (2010).
[18] E. Kitamura, J. J. Blow, and T. U. Tanaka, Cell **125**, 1308 (2006).
[19] K. L. Friedman, B. J. Brewer, and W. L. Fangman, Genes Cells **2**, 667 (1997).
[20] N. Rhind, S. C.-H. Yang, and J. Bechhoefer, Chromosome Res. **18**, 35 (2010).
[21] E. Shor, C. L. Warren, J. Tietjen, Z. Hou, U. Müller, I. Alborelli, F. H. Gohard, A. I. Yemm, L. Borisov, J. R. Broach, M. Weinreich, C. A. Nieduszynski, A. Z. Ansari, and C. A. Fox, PLoS Genet. **5**, e1000755 (2009).
[22] A. Kolmogorov, Izv. Akad. Nauk. SSSR **1**, 335 (1937).
[23] Y. A. Andrienko, N. V. Brilliantov, and P. L. Krapivsky, Phys. Rev. A **45**, 2263 (1992).
[24] A. Al-Mahboob, Y. Fujikawa, J. T. Sadowski, T. Hashizume, and T. Sakurai, Phys. Rev. B **82**, 235421 (2010).
[25] N. Yabuki, H. Terashima, and K. Kitada, Genes Cells **7**, 781 (2002).
[26] D. M. Gilbert, Nat. Rev. Genet. **11**, 673 (2010).
[27] C. A. Müller and C. A. Nieduszynski, Genome Res. (2012), doi: 10.1101/gr.139477.112 (2012).
[28] K. Woodfine, H. Fiegler, D. Beare, J. Collins, O. McCann, B. Young, S. Debernardi, R. Mott, I. Dunham, and N. Carter, Hum. Mol. Genet. **13**, 191 (2004).
[29] D. Schubeler, D. Scalzo, C. Kooperberg, B. van Steensel, J. Delrow, and M. Groudine, Nat. Genet. **32**, 438 (2002).
[30] T. Ryba, D. Battaglia, B. D. Pope, I. Hiratani, and D. M. Gilbert, Nat. Protoc. **6**, 870 (2011).
[31] E. Yaffe, S. Farkash-Amar, A. Polten, Z. Yakhini, A. Tanay, and I. Simon, PloS Genet. **6**, e1001011 (2010).
[32] A. Koren, I. Soifer, and N. Barkai, Genome Res. **20**, 781 (2010).
[33] A. Koren, H.-J. Tsai, I. Tirosh, L. S. Burrack, N. Barkai, and J. Berman, PloS Genet. **6**, e1001068 (2010).
[34] M. Ma, O. Hyrien, and A. Goldar, Nucleic Acids Res. **40**, 2010 (2012).
[35] A. Goldar, M.-C. Marsolier-Kergoat, and O. Hyrien, PloS One **4**, 1 (2009).
[36] M. K. Raghuraman and B. J. Brewer, Chromosome Res. **18**, 19 (2010).
[37] M. Anglana, F. Apiou, A. Bensimon, and M. Debatisse, Cell **114**, 385 (2003).
[38] J. Bechhoefer and B. Marshall, Phys. Rev. Lett. **98**, 098105 (2007).
[39] D. Fachinetti, R. Bermejo, A. Cocito, S. Minardi, Y. Katou, Y. Kanoh, K. Shirahige, A. Azvolinsky, V. A. Zakian, and M. Foiani, Mol. Cell **39**, 595 (2010).
[40] J. J. Blow, X. Q. Ge, and D. A. Jackson, Trends Biochem. Sci. **36**, 405 (2011).
[41] A. Letessier, G. A. Millot, S. Koundrioukoff, A.-M. Lachages, N. Vogt, R. S. Hansen, B. Malfoy, O. Brison, and M. Debatisse, Nature (London) **470**, 120 (2011).
[42] P. Patel, B. Arcangioli, S. Baker, A. Bensimon, and N. Rhind, Mol. Biol. Cell **17**, 308 (2006).
[43] J. Blow and A. Dutta, Nat. Rev. Mol. Cell Biol. **6**, 476 (2005).
[44] C. Conti, B. Sacca, J. Herrick, C. Lalou, Y. Pommier, and A. Bensimon, Mol. Biol. Cell **18**, 3059 (2007).
[45] S. Tuduri, H. Tourrière, and P. Pasero, Chromosome Res. **18**, 91 (2010).
[46] C. Rivin and W. Fangman, J. Cell Biol. **85**, 108 (1980).
[47] D. J. Smith and I. Whitehouse, Nature (London) **483**, 434 (2012).
[48] Parameter values used were $q_1 = 0.96$, $q_2 = 0.88$, $\tau = 4$, $\Delta t = 10$, and $v = 1.6$ kb/min, estimated as described in [4].
[49] M. Lundgren, A. Andersson, L. Chen, P. Nilsson, and R. Bernander, Proc. Natl. Acad. Sci. USA **101**, 7046 (2004).
[50] C. Norais, M. Hawkins, A. L. Hartman, J. A. Eisen, H. Myllykallio, and T. Allers, PloS Genet. **3**, 729 (2007).
[51] I. G. Duggin, N. Dubarry, and S. D. Bell, EMBO J. **30**, 145 (2011).
[52] C. C. Siow, S. R. Nieduszynska, C. A. Müller, and C. A. Nieduszynski, Nucleic Acids Res. **40**, D682 (2011).
[53] Parameter values used were $q_1 = 0.68$, $q_2 = 0.73$, and $v = 1.6$ kb/min, estimated in [4].
[54] X. Wang, C. Lesterlin, R. Reyes-Lamothe, G. Ball, and D. J. Sherratt, Proc. Natl. Acad. Sci. USA **108**, E243 (2011).