Modeling Inhomogeneous DNA Replication Kinetics

Michel G. Gauthier[1], **Paolo Norio**[2,3], **John Bechhoefer**[1]*

1 Department of Physics, Simon Fraser University, Burnaby, British Columbia, Canada, 2 Department of Oncology, Montefiore Medical Center, Moses Division, Bronx, New York, United States of America, 3 Department of Medicine, Albert Einstein College of Medicine, Bronx, New York, United States of America

Abstract

In eukaryotic organisms, DNA replication is initiated at a series of chromosomal locations called origins, where replication forks are assembled proceeding bidirectionally to replicate the genome. The distribution and firing rate of these origins, in conjunction with the velocity at which forks progress, dictate the program of the replication process. Previous attempts at modeling DNA replication in eukaryotes have focused on cases where the firing rate and the velocity of replication forks are homogeneous, or uniform, across the genome. However, it is now known that there are large variations in origin activity along the genome and variations in fork velocities can also take place. Here, we generalize previous approaches to modeling replication, to allow for arbitrary spatial variation of initiation rates and fork velocities. We derive rate equations for left- and right-moving forks and for replication probability over time that can be solved numerically to obtain the mean-field replication program. This method accurately reproduces the results of DNA replication simulation. We also successfully adapted our approach to the inverse problem of fitting measurements of DNA replication performed on single DNA molecules. Since such measurements are performed on specified portion of the genome, the examined DNA molecules may be replicated by forks that originate either within the studied molecule or outside of it. This problem was solved by using an effective flux of incoming replication forks at the model boundaries to represent the origin activity outside the studied region. Using this approach, we show that reliable inferences can be made about the replication of specific portions of the genome even if the amount of data that can be obtained from single-molecule experiments is generally limited.

Citation: Gauthier MG, Norio P, Bechhoefer J (2012) Modeling Inhomogeneous DNA Replication Kinetics. PLoS ONE 7(3): e32053. doi:10.1371/journal.pone.0032053

Editor: Jerome Mathe, Université d'Evry val d'Essonne, France

Received October 19, 2011; **Accepted** January 20, 2012; **Published** March 7, 2012

Funding: Michel G. Gauthier and John Bechhoefer were supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada and by the Human Frontier Science Program (HFSP). Paolo Norio was supported by National Institutes of Health grant R01GM080606. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: johnb@sfu.ca

Introduction

Cells must accurately duplicate their DNA content at every cell cycle. Depending on the organism, the process of DNA replication can initiate at one or multiple sites called origins of replication. The DNA is copied by a pair of oppositely moving replication forks that emerge from each origin. These forks actively replicate the genome away from the origin until they encounter another replication fork. DNA replication can thus be modeled as a series of nucleations, growth (perhaps including fork stalls and rescues [1,2]), and coalescences occurring in an asynchronous parallel way until the whole genome is copied [3,4] (Fig. 1).

The complexity of the replication process traces back to the observation that the initiation program can be inhomogeneous in both space and time (see [5–11] for examples). Spatially inhomogeneous replication firing can be caused by a variety of factors such as an inhomogeneous distribution of pre-replication complexes or their uneven activation during the S phase. This is believed to be caused by factors such as the primary sequence of DNA, the presence of transcription factor binding sites, the chromatin organization of the DNA template and by gene expression [5,12,13]. The variability of origin initiation times, on the other hand, can result from the stochastic recruitment of replication initiation factors and the level of checkpoint activity [14–16]. As a consequence of such stochastic initiation, replication origins can also be passively replicated by forks coming from neighboring origins. In summary, modeling DNA replication is challenging because the probability of initiation of an origin varies along the genome, the moment at which an origin fires is stochastic, and origins do not systematically fire at each cell cycle.

DNA replication modeling is also challenged by the lack of direct observations. Experimental techniques using immunofluorescent labels to observe the DNA synthesis provide only snapshots of the replication kinetics [17]. The modeling approach presented in this paper can be used to reveal the detailed replication program responsible for producing these snapshots (initiation rates, fork speeds, stalling events, etc).

Over the last decade, our group has developed an analytic approach to modeling DNA replication kinetics [3,4,18–25]. The approach is based on a formalism inspired by the Kolmogorov-Johnson-Mehl-Avrami (KJMA) theory of phase-transition kinetics in one spatial dimension [26–30]. In general, this approach has assumed that there was no significant spatial variation along the genome in the parameters characterizing replication. (Except for Ref. [18] in which we looked at replication in budding yeast, where origins have fixed locations. Reference [18] turns out to be somewhat different from the present case, where origin initiation occurs in extended zones that then show variation along the genome.) In particular, we assumed that origin initiation rates and the rate of DNA synthesis (fork progression velocities) were spatially uniform. Temporal variations, however, were included, and their effects can be important [18,20,22–24]. Because our

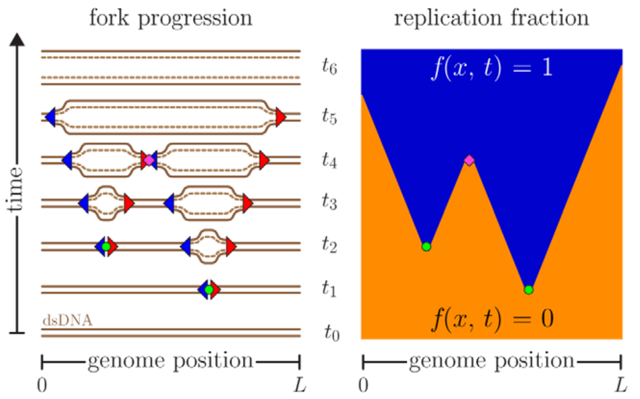


Figure 1. Space-time representation of the replication kinetics. The left-hand side shows the original (solid lines) and new synthesized (dashed lines) DNA while replications forks (triangles) are moving. In this example, the forks originate from two origins (circles) that are initiated at times t_1 and t_2. The forks move at a constant speed until they coalesce with another fork (diamond at t_4) or reach the ends of the molecule of length L (around t_5 and t_6). The right-hand side presents the space-time replication fraction $f(x,t)$, where x is the position along the genome, of the same replication cycle. Orange and blue areas represent unreplicated ($f=0$) and replicated DNA ($f=1$), respectively.
doi:10.1371/journal.pone.0032053.g001

approach gives analytical solutions for the evolution of experimentally measurable quantities such replication progress, fork densities, domain densities, and the like, it is particularly well suited for fitting to experimental data [18,20]. This offers an advantage compared to other approaches based on lengthy Monte Carlo simulations [31–35] because it requires far less computational power to fit experimental data.

In this paper, we generalize our analytic approach to the case where initiation rates and fork velocities may vary in both space and time. We derive simple rate equations that can be solved numerically to obtain the mean-field space-time replication kinetics. We find the average fork densities in both directions, everywhere along the genome and at any moment during the synthesis (S) phase of the cell cycle. This technique can be used to analyze experimental data from molecular combing [36,37] and microarrays [38–40]. In addition, since our approach allows us to determine quantities involving DNA replication initiation, progression and termination (e.g., coalescence probability profiles, replication time distributions, etc.), it is particularly suitable for fitting results obtained from experiments based on the single-molecule analysis of replicated DNA (SMARD) where molecules at all stages of DNA replication are considered and the steady-state distribution of replication forks can be determined within a specific portion of the genome [7]. On the other hand, the mean-field assumption assumes that the cell-to-cell variations in parameters relevant to replication are small. It also does not give the statistical variation expected from an analysis of a finite number of cells, even when all cells are identical. Both of these limitations can be addressed by Monte Carlo simulations, which should be seen as complementary to the present approach.

Methods

Simulating DNA replication

Although the goal of this paper is to be able to calculate the average replication kinetics without recourse to numerical simulations, we shall use simulations here to test our model solutions and, more extensively, to test our fitting procedures. As illustrated in Fig. 1, we model DNA replication using a series of origins from which a pair of replication forks emerge to bidirectionally duplicate the DNA. These forks move away from the initiation site until they coalesce with another fork or reach the end of the molecule. At this level of description, only the rate at which forks are initiated, $I(x,t)$, as well as their propagation speed, $v(x,t)$, is needed in order to simulate the process of DNA replication. We previously used a Monte Carlo simulation to study the case in which origin initiation rates and fork progression are spatially homogenous along the genome, i.e., $I(x,t)=I$ and $v(x,t)=v$. Such processes are described in detail in Ref. [3]. However, experimental observations indicate that initiation rates can vary in both space and time along the genome and that the speed of replication forks is not necessarily uniform. Hence, the Monte Carlo simulations must be modified to model these inhomogeneous factors. In addition, since in mammalian cells initiation events frequently appear scattered across large genomic regions (rather than being limited to the precise DNA sequences), we included in our simulations the presence of initiation zones. We chose Gaussian profiles for the zones. Although the form of such zones is not clear experimentally, our formalism can work with zones that have an arbitrary initiation-rate profile along the genome (x-axis).

As a test case for our new model, we simulated the replication of a genomic region of 1000 kb containing two Gaussian initiation zones of similar size (50 kb), as indicated in Fig. 2. Each zone is assumed to contain origin that fire at different times during the S phase and therefore referred to as "early" and "late." In Fig. 2, the early zone is centered at 200 kb and is active at all times. The late zone, located at 800 kb, on the other hand, is active only for $t \gtrsim 5000$ sec. The late zone is also assumed to be 10 times more efficient than the early one (1.6×10^{-6} initiations/kb/sec at the peak of the early zone). The initiation rate $I(x,t)$ indicates the average number of initiations that occur at (x,t) per length of unreplicated DNA per unit of time. This definition is motivated by the observation that each portion of the genome replicates only once per S phase. For this specific example, we set the fork velocity profile, $v(x,t)$ to a constant value of 0.04 kb/sec. The simulation parameters chosen here are typical of replication in somatic cells in

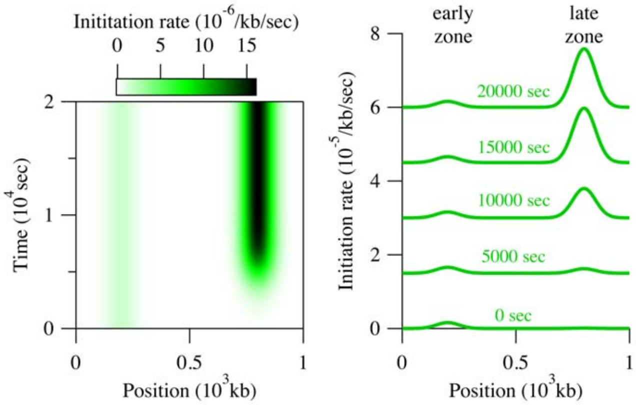


Figure 2. Initiation profile $I(x,t)$ used to produce the results presented in Fig. 3. The left-hand side is a density plot of the initiation rate, while the right-hand side shows $I(x)$ at various time points. (For clarity, each curve is offset by 15×10^{-6}/kb/sec from the previous one.) The initiation pattern is composed of two Gaussian initiation zones at 200 and 800 kb. The first, or "early,"zone is constant throughout time, while the second, more efficient, "late" zone is turned on at 5000 sec.
doi:10.1371/journal.pone.0032053.g002

mammalian organisms [41]. Our method can easily include variable velocities, including fork blocks due to DNA damage. (Our approach allows both I and v to be space-time dependent but the results of our test case are easier to interpret when there is only one inhomogeneous contribution to the replication kinetics.) However, experimental results indicate that the effects of the inhomogeneity of $I(x,t)$ are much more important than the effects of the inhomogeneity of $v(x,t)$ (see below and Demczuk *et al.*, unpublished). For simplicity, we used periodic boundary conditions (PBCs) for the fork propagation in our simulations (forks reaching a boundary are re-inserted at the other boundary). Therefore, it is formally equivalent to a circular chromosome (e.g., as in bacteria). Of course, whole-chromosome simulations of eukaryotic chromosomes would not use periodic boundary conditions and would take into account the specific (low) initiation rates found in telomeres [38].

Simulation results for our model system for 1 and 1000 cell cycles are presented in Figs. 3 a and 3 b, respectively. Each figure is divided into five parts. Part I shows the replication fraction $f(x,t)$. For the one-cycle simulation, the value of f is either 0 or 1 for unreplicated and replicated DNA. For a finite number of simulations, as in Fig. 3 b–I, the value of $f(x,t)$ is the average value observed throughout the ensemble of simulated cycles $(0 \leq f(x,t) \leq 1)$. The fraction $f(x,t)$ thus gives the probability that a specific section of the genome located at x is replicated a time t. Parts II and III present the left- and right-moving replication forks. Only the trajectories of the forks are displayed for the case of one simulated cycle, while the average observed fork densities are reported for the case of many cycles. We refer to the fork densities presented Fig. 3 b–II and 3 b–III as $\rho_{\pm}(x,t)$, where the ± sign represents the right- and left-moving forks, respectively. These densities equal the number of forks moving in a specific direction per kb at (x,t). Finally, parts IV and V show where and when initiations and coalescences occur. In b–IV and b–V, these events are represented using probability density functions, $\phi_{\mathrm{i}}(x,t)$ and $\phi_{\mathrm{c}}(x,t)$, for initiations and coalescences, respectively. These densities are expressed in units of 1/kb/sec and are normalized so that $\int_0^\infty \int_0^L \phi_{\mathrm{i,c}}(x,t)\,dx\,dt = 1$.

Our simulations give detailed information about the replication process and its statistics. Typical quantities of interest that we study include the distribution of whole-genome replication times, the average replication times of different regions, and the average number of initiations and coalescences (as well as their space-time distributions). However, while simulations based on a known scenario can reproduce any experimentally obtainable statistic, the calculation times are long. To fit unknown parameters to a set of

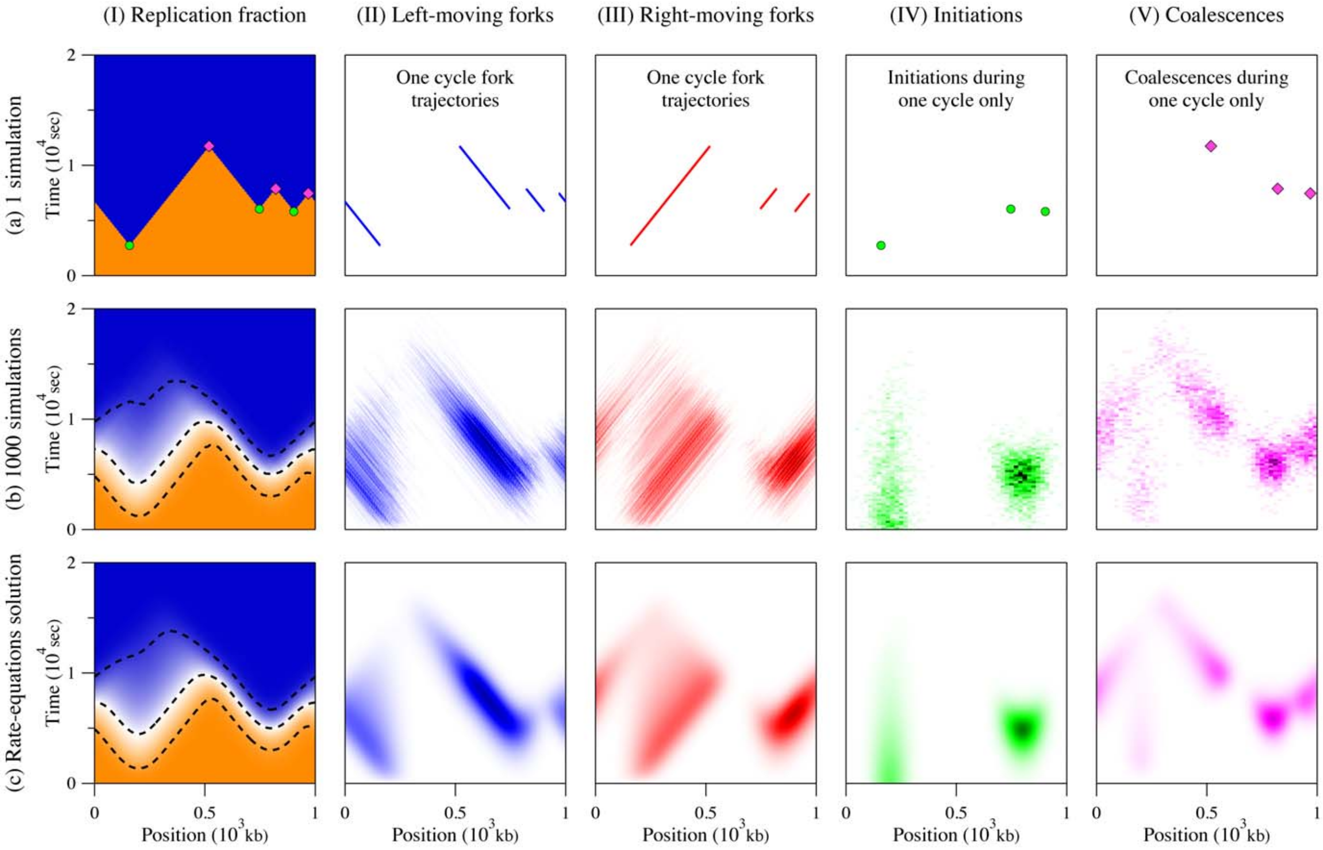


Figure 3. Comparison between one simulated replication cycle (a), 1000 simulation cycles (b), and our rate-equation solution (c). In graph (I), the color scale goes from 0 (orange) to 1 (blue); in graphs (II) and (III), it goes from 0 (white) to 0.01/kb (black); in graphs (IV) and (V), it goes from 0 (white) to 1.5×10^{-6}/kb/sec (black). In all cases, we used the initiation function $I(x,t)$ presented in Fig. 2 and the fork velocity $v(x,t) = 0.04$ kb/sec. The genome size is 1000 kb, with periodic boundary conditions. Column I compares the replication fraction $f(x,t)$ in the three cases. The dashed lines in b–I and c–I show the 10%, 50% and 90% contour curves. Columns II and III present the fork densities $\rho_{\pm}(x,t)$. Fork densities are expressed in forks/kb in (b) and (c) while trajectories only are shown for the single cell cycle in (a). Columns IV and V present the space-time probability density functions of observing an initiation, $\phi_{\mathrm{i}}(x,t)$, or a coalescence, $\phi_{\mathrm{c}}(x,t)$, respectively. Part (a) shows where and when initiations and coalescences from one cycle occurred while parts (b) and (c) represent probability densities in 1/kb/sec.
doi:10.1371/journal.pone.0032053.g003

experimental data would require large computational resources. This difficulty motivates the analytic methods presented in the next section. Although they use numerical methods to solve differential equations, they are orders of magnitude faster than simulation-based approaches.

Rate-equation approach

As mentioned above, we have developed a theoretical approach that can be substituted for numerical simulations in order to speed up the analysis of a given replication scenario when one is interested in the average replication kinetics. As we will show, integrating our rate-equations system also involves numerical steps, but our approach is still considerably faster than simulation-based models. Moreover, our method directly gives the mean-field kinetics of replicating DNA. This solution is equivalent to the simulation results in the limit where an infinite number of simulations is performed. (Compare Fig. 3 b to Fig. 3 c). In this sense, our technique provides the exact average replication program but does not give information about the cell-to-cell variability of the process, which can be obtained from simulations. Simulations are thus complementary to the present mean-field calculation method.

In this section, we introduce an analytical formalism to model the creation, propagation and annihilation of replication forks during DNA synthesis. To proceed, we derive a set of coupled differential equations that describe the change in the replication fork populations as a function of both the position along the genome (x) and the time since the beginning of S phase (t). As before, we define $f(x,t)$ to be the probability that a given position of the genome x is replicated at time t, while $\rho_{\pm}(x,t)$ represents the right- and left-moving fork densities.

Modeling the replication kinetics using rate-equations. We describe the space-time evolution of the average replication fraction $f(x,t)$ as well as both average fork densities $\rho_{\pm}(x,t)$, assuming that the creation and propagation dynamics of the forks are inhomogeneous, i.e., that the initiation function $I(x,t)$ and the fork speeds $v_{\pm}(x,t)$ can vary in space and time. (Again, the ± signs refer to the direction of propagation of the forks.)

The first equation of our set gives the rate of change of the probability that a given location, x, is replicated at time t,

$$\frac{\partial f}{\partial t} = (v_{-}\rho_{-}) + (v_{+}\rho_{+}), \tag{1}$$

which is simply given by the product of local fork densities times the rate at which a given fork synthesizes DNA.

The rate of change of the fork densities can be expressed in the form of a "transport" equation,

$$\frac{\partial \rho_{\pm}}{\partial t} \pm \frac{\partial (v_{\pm}\rho_{\pm})}{\partial x} = I(1-f) - \frac{(v_{-}+v_{+})\rho_{-}\rho_{+}}{1-f}, \tag{2}$$

with a "source" and a "sink" term on the right-hand side. The source term, $I(1-f)$, represents the initiation of new forks at a rate $I(x,t)$ rescaled by the probability that the genome is not already replicated at that position, $1-f(x,t)$. The sink term represents the annihilation rate of forks as they coalesce with oppositely moving forks. The coalescence rate is proportional to the two local populations of forks and the relative speed at which these forks are merging. That rate must be normalized by the probability of not being replicated, $1-f(x,t)$. The ± sign on the left-hand side of Eq. 2 arises because both left- and right-moving forks are assigned positive velocities. An expression similar to Eq. 2 has been used to model the growth of crystal lamella [42].

Given a replication scenario for $I(x,t)$ and $v_{\pm}(x,t)$, Eqs. 1 and 2 can be numerically integrated to obtain $f(x,t)$ and $\rho_{\pm}(x,t)$. We solved our set of equations for the same conditions as used for the simulations presented in Fig. 3 b (i.e., $I(x,t)$ given by Fig. 2 and $v_{\pm}(x,t) = 0.04$ kb/sec). We explicitly integrated our equations using $dx = 0.\bar{3}$ kb and $dt = 8.\bar{3}$ sec (we need $dx/dt \geq v$ to adequately solve this system). We also used PBCs to solve our equations, which means we used $f(0,t) = f(L,t)$ and $\rho_{\pm}(0,t) = \rho_{\pm}(L,t)$ for all t. Using $f(x,0) = 0$ and $\rho_{\pm}(x,0) = 0$ for all x as initial conditions, the solution presented in Fig. 3 c agrees with the simulation results of Fig. 3 b within statistical limits. Parts I to III are directly obtained from the solution of our three rate-equations. The densities of parts IV and V are, on the other hand, proportional to the source and sink terms of Eq. 2 , respectively. Hence,

$$\phi_{\mathrm{i}} = \frac{I(1-f)}{N_{\mathrm{i}}}, \tag{3}$$

where

$$N_{\mathrm{i}} = \int_{0}^{\infty}\int_{0}^{L} I(1-f)\,dx\,dt, \tag{4}$$

is the average number of initiations per replication cycle. Similarly,

$$\phi_{\mathrm{c}} = \frac{(v_{-}+v_{+})\rho_{-}\rho_{+}}{N_{\mathrm{c}}(1-f)}, \tag{5}$$

where the average number of coalescences per cycle is given by

$$N_{\mathrm{c}} = \int_{0}^{\infty}\int_{0}^{L} \frac{(v_{-}+v_{+})\rho_{-}\rho_{+}}{(1-f)}\,dx\,dt. \tag{6}$$

The results of our numerical integrations are $N_{\mathrm{i}} = 3.251$ and $N_{\mathrm{c}} = 3.244$. For a finite-size system with periodic boundary conditions such as the model presented in this paper, we must have $N_{\mathrm{c}} = N_{\mathrm{i}}$. The 0.2% difference between our calculation results is simply due to round-off errors. Our calculation also matches the average number of 3.25 ± 0.04 initiations observed during our 1000 simulations. Finally, note that our model can also be solved using non-periodic boundary conditions in order to study replication of linear DNA. In such a case, the numbers of initiations N_{i} and coalescences N_{c} are still given by Eqs. 4 and 6, but we expect $N_{\mathrm{i}} = N_{\mathrm{c}} + 1$.

Start-time distributions. The stochasticity of the replication process modeled here implies that the start and end of S phase (defined by the first origin initiation and the last fork coalescence) occur at different times each single cycle. As illustrated in Fig. 3 a–I, the simulation starts at $t = 0$, but the actual duplication of the DNA does not start before $t \approx 3000$ sec. In other words, there is a distribution of replication start times (marked by the first initiation) and also a distribution of end times (marked by the last coalescence).

Our rate equations can be used to calculated the probability that replication has started, $P_{\mathrm{s}}(t)$, as a function of time, which corresponds to the probability that at least one initiation occurs during the time interval $[0,t]$. We calculate this probability in terms of a related quantity, $N_{\mathrm{expt}}(t)$, which is the number of initiations that are expected to happen in $[0,t]$, assuming that there were no initiations prior to $t = 0$. We write

$$N_{\mathrm{expt}}(t)=\int_0^t\int_0^L I(x,t')\,dx\,dt', \tag{7}$$

where L is the genome length. Consequently, the probability that at least one initiation occurred prior to time t is given by

$$\begin{aligned}P_{\mathrm{s}}(t)&=1-e^{-N_{\mathrm{expt}}(t)}\\&=1-\exp\left\{-\int_0^t\int_0^L I(x,t')\,dx\,dt'\right\},\end{aligned} \tag{8}$$

and the replication starting time distribution is simply given by $\phi_{\mathrm{s}}(t)=dP_{\mathrm{s}}(t)/dt$. Figure 4 compares the calculated starting time distribution with simulation results.

Equation 8 is valid for any molecule of length L, whether periodic or non-periodic boundary conditions are considered. However, Eq. 8 must be modified if one is studying a finite-size fragment that is part of a larger molecule. A fragment thus corresponds to a finite-length linear molecule without PBCs but with a flux of forks at its boundaries (these forks were previously initiated elsewhere outside the fragment region). In order to calculate the starting probability of such a fragment, we first define the notion of *directional replication fractions*, $f_\pm(x,t)$, which are the probabilities that the location x has been replicated by a right- or a left-moving fork. These replication fractions are obtained from

$$\frac{\partial f_\pm}{\partial t}=v_\pm\rho_\pm \tag{9}$$

and can be calculated as a by-product of the numerical integration of Eq. 1 . Thus, for a fragment that begins at x_- and ends at x_+, the probability that replication has started is given by

$$P_{\mathrm{s}}(t)=1-e^{-N_{\mathrm{expt}}(t)}[1-f_+(x_-,t)][1-f_-(x_+,t)], \tag{10}$$

where

$$N_{\mathrm{expt}}(t)=-\int_0^t\int_{x_-}^{x_+} I(x,t')\,dx\,dt'. \tag{11}$$

Equation 10 says that the probability that replication has not started along a given molecule is the product of the probability that no initiation already occurred within the molecule times the probability that no fork came across the molecule boundaries.

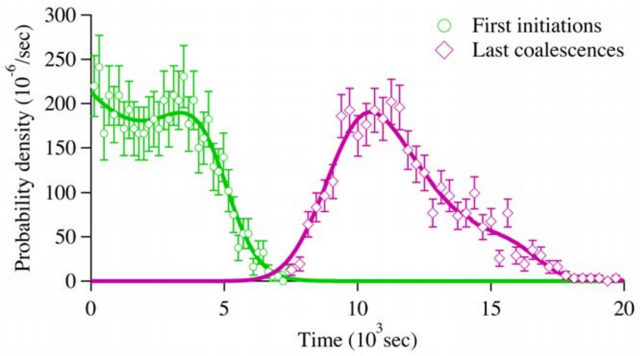


Figure 4. Replication starting and ending times density functions, $\phi_{\mathrm{s}}(t)$ and $\phi_{\mathrm{e}}(t)$, for our model system. Symbols were obtained from simulations, while solid lines were calculated from the solution of our rate-equations.
doi:10.1371/journal.pone.0032053.g004

End-time distributions. Another useful quantity is the probability that replication has ended at time t, $P_{\mathrm{e}}(t)$. This quantity is of great interest because it tells us when the replication is over. It could therefore be used to study the duplication time of the genome as a function of the replication scenarios. In general, we cannot derive an analytical solution of our nonlinear rate-equations system and, consequently, we cannot derive a formula for $P_{\mathrm{e}}(t)$ as we did for the starting-time distribution. Nonetheless, we can use our knowledge of the replication fork density to estimate $P_{\mathrm{e}}(t)$ as the probability that there is no fork along the genome. For a periodic system, where the number of right-moving forks is always equal to the left-moving forks, we have

$$P_{\mathrm{e}}(t)\approx\tilde{P}_{\mathrm{e}}(t)=\exp\left\{-\int_0^L\frac{\rho_+(x,t)}{f(x,t)}dx\right\}, \tag{12}$$

where we have assumed the number of forks at time t to be given by a Poisson distribution (an equivalent estimate for $P_e(t)$ in a system with PBCs is obtained by replacing ρ_+ $(x;\ t)$ by ρ_- $(x,\ t)$ in Eq. 12). The tilde notation used in Eq. 12 denotes the fact that $\tilde{P}_{\mathrm{e}}$ is an approximation of $P_{\mathrm{e}}(t)$. The density is normalized by the probability of being replicated. Figure 4 also compares the end-time distribution function, $\phi_{\mathrm{e}}(t)=d\tilde{P}_{\mathrm{e}}(t)/dt$, with simulation results. Note that we can replace $f(x,t)$ by $1-f(x,t)$ in Eq. 12 to get an estimate of $1-P_{\mathrm{s}}(t)$. We expect the approximation that fork distributions are Poisson to be accurate at the beginning and end of S phase, where the number of forks is small, but to be less so in mid-S phase, where there are more forks. For the model explored here, the maximum difference between the calculated and the simulated values of the ending probability is ≈ 0.03. In Supporting Information S1, we solve exactly the case of a uniform initiation profile and show that the error of our approximation of $P_{\mathrm{e}}(t)$, when compared to the exact solution, decreases as the number of initiations increases (i.e., as $L\sqrt{I/v}$ increases).

In the case of a non-periodic system, the lack of forks that move in one direction over a certain range $[x_-,x_+]$ does not imply that the whole range is replicated. Therefore, we must modify Eq. 12 so as to obtain the end-time distribution of finite-size systems without periodic boundary conditions (e.g., finite-length linear DNA or a section of a larger molecule). The probability that a DNA fragment located between x_- and x_+ is fully replicated is given by

$$\tilde{P}_{\mathrm{e}}(t)\approx f(x_-,t)\exp\left\{-\int_{x_-}^{x_+}\frac{\rho_+(x,t)}{f(x,t)}dx\right\}. \tag{13}$$

Equation 13 asserts that the replication of a molecule without PBCs has finished if no right-moving forks are observed and if the left boundary is replicated (or vice versa for left-moving forks). As mentioned above, an equivalent estimate for $P_e(t)$ without PBCs is obtained if we substitute the pre-factor f(x_-, t) by f(x_+, t) and use ρ_- (x, t) instead of ρ_+(x, t).

Boundary fork injection. The previous sections presented how our model can be used to study replication of molecules with and without PBCs. Deriving Eqs. 10 and 13, we even demonstrated how to calculate the probability that a sub-section of the modeled systems has started or ended replicating. Here we now show how we can adapt the boundary condition so they act as sources of forks in order to account for initiations that occur outside the modeled DNA segment. These forks mimic initiations occurring outside $[0,L]$. The simplest case would be to have a source term that is equivalent to a semi-infinite region where the initiation rate and the fork velocity are constant. In such a case, the density of forks at the boundaries is simply

$$\rho_+(x=0,t)=I_+\,t\,[1-f(x=0,t)], \tag{14}$$

$$\rho_-(x=L,t)=I_-\,t\,[1-f(x=L,t)], \tag{15}$$

where $I_\pm$ are the constant initiation rates outside the modeled regions (I_+ for right-moving forks coming from the $x<0$ region and I_- for left-moving forks initiated at $x>L$). The derivation for this boundary condition is presented in Supporting Information S1 and the Figure S1.

Stochastic fork progression. Our calculation method can also be adapted to model the impact of DNA damage on replication kinetics. Even in normal, healthy cells, there are a large number of DNA "defects" where forks slow, or even stop. Such damage usually affects only one of the two DNA strands. These single-strand lesions are characterized by base oxidation caused by reactive oxygen species or by base misincorporation due to a copying error during DNA replication. In more serious but rarer cases, defects involve both DNA strands. Examples of such double-strand defects include DNA crosslinking induced by ionizing radiation or double-strand breaks that result from a failed repair to single-strand damage. Double-strand damage is more dangerous because its repair can lead to rearrangements of the genome and even contribute to the development of cancer [43]. Depending on their density and on the repair mechanisms involved, DNA damage can have a strong impact on the replication kinetics. The slow down or stalling of forks at defects gives more time to fire to origins that would otherwise have been passively replicated [44]. Also, fork stalls trigger local and global checkpoint signals that can affect the progression of forks and the firing rate of new origins elsewhere along the genome [11].

If replication speed changes predictably along the genome, one can simply define an appropriate velocity profile $v(x,t)$. However, fork progression can also be affected in a more stochastic way in the presence of DNA damage. When they encounter such defects, replication forks are stalled for a given period of time until repaired. The repair time depends on the nature of the defects and can either be finite or infinite (i.e., not repaired during the current S phase). In the infinite-repair-time case, the replication of the DNA on the other side of the defect must come from the oppositely moving fork. Such a stochastic blocking/unblocking mechanism can be added to our mathematical framework by modifying our expression for $\partial\rho_\pm/\partial t$ to

$$\begin{aligned}\frac{\partial\rho_\pm}{\partial t}\pm\frac{\partial(v_\pm\rho_\pm)}{\partial x}&=I(1-f)-\frac{v_\pm\rho_\pm}{\delta}+\frac{\rho'_\pm}{\tau}\\&-\frac{(v_-+v_+)\rho_-\rho_+}{1-f}-\frac{v_\pm\rho_\pm\rho'_\mp}{1-f},\end{aligned} \tag{16}$$

where the five terms on the right hand side are

1. the initiation rate of new forks, as we had in Eq. 2 ;
2. the stall rate of the moving forks, assuming that the average spacing between defects is given by δ;
3. the repair rate of the stalled forks, denoted $\rho'_\pm$, with the average repair time given by τ;
4. the coalescence rate between moving forks, per Eq. 2 ;
5. the coalescence rate of moving forks that collide with stalled forks.

The densities of stalled forks can be obtained by adding two differential equations to our set. These new equations are used to describe the rate of change of the densities of forks that are stalled at DNA lesions as

$$\frac{\partial\rho'_\pm}{\partial t}=\frac{v_\pm\rho_\pm}{\delta}-\frac{\rho'_\pm}{\tau}-\frac{v_\pm\rho_\pm\rho'_\mp}{1-f}, \tag{17}$$

where the three terms represent stall, repair, and coalescence rates. There is no $\partial/\partial x$ term on the left-hand side of Eq. 17 because stalled forks are assumed to be fixed in space. A simplified version of this fork-stall model, neglecting spatial inhomogeneity, was the subject of a previous publication [19].

Results

Analyzing experimental data

An obvious application of our analysis would be to reproduce results from experiments based on microarrays [38]. Microarrays provide genome-wide average replication profile as a function of time (derived from the overall molecule replication fraction), which ideally corresponds to the replication fraction $f(x,t)$ obtained from our rate equations (see [39,40] for examples). Of course, real microarray experiments are not ideal, and issues such as the spatial resolution of the array or the cell-cycle asynchrony of populations should be kept in mind when analyzing the data. In a future contribution, we shall discuss how to reproduce such time-course results. Here, we demonstrate the versatility of our modeling technique by adapting it to the study of a more subtle type of data that has recently been obtained via single molecule analysis of replicated DNA (SMARD), a method developed by Norio *et al.* [7]. The modeling and fitting procedures presented in this paper were used to analyze a large SMARD data set obtained from mice bone marrow cells (Demczuk *et al.*, unpublished). One feature of such experiments is that the data are obtained from an asynchronous population of cells (i.e., the starting time of each cell in the population is random, drawn from a uniform distribution). Unlike microarrays, SMARD also allows one to determine the steady state distribution of replication forks, as well as the location of initiation events and fork collisions (in addition to the temporal order of replication for a specific portion of the genome). This additional information can be used to determine more precisely the level of origin activity across the genomic region analyzed. We shall need to adapt our model to make predictions for such a case.

Simulating a SMARD data set. The goal of the current section is to adapt our calculation approach to the analysis of an actual experimental setup, the SMARD experiment. The first step towards such a goal is to be able to simulate the data collected during this experiment.

The SMARD procedure is presented in detail in Ref. [7]. Here, we give a brief summary. In a population of asynchronously growing cells, one supplements the normal nucleotides used to synthesize DNA by two different types of halogenated nucleotides that are then conjugated to fluorescent antibodies. For convenience, we shall refer to them as red and green labels. (The first label is red; the second is green). Since cells are replicate asynchronously, the labeling switch can occur at any time relative to the cell cycle for a particular cell. (In particular, the switch will often occur when the cell is not in S phase.) Figure 5 depicts the labeling procedure when the transition happens during the replication process. Part (a) compares the labeling timeline with the replication space-time diagram, while part (b) shows the DNA molecule one would observed after such labeling. As shown in Fig. 5 b, the positions where labels are changing indicate the

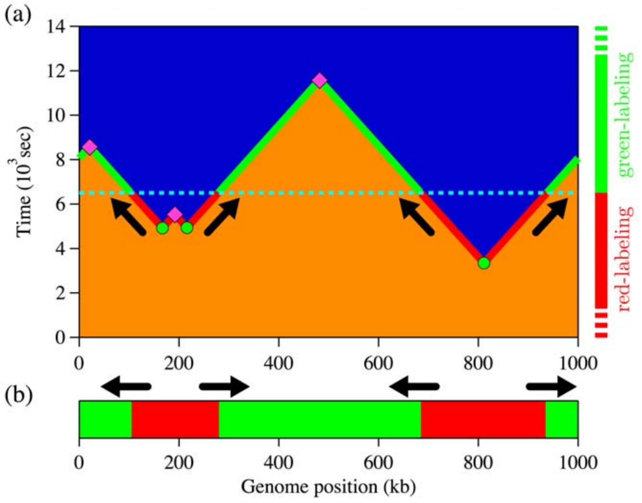


Figure 5. SMARD labeling procedure. (a) Example of a replication space-time profile and the corresponding SMARD labeling procedure. As before, blue sections indicate replicated DNA while orange sections represent unreplicated DNA. Circles denote fired origins, while diamonds indicate coalescences of replication forks. Periodic boundary conditions were used (circular genome). The dashed line at time $t = 6500$ sec indicates the end of the first labeling period (red) and the beginning of the second (green) one. Arrows indicates the fork propagation directions at the labeling transition time. The labeling timeline on the right side and the solid line on the space-time profile illustrate the labeling process to produce the molecule example presented in (b). (b) Example of a molecule extracted from the simulation presented in (a). Red sections were replicated during the red pulse (before $t = 6500$ sec), while green sections were replicated later. To obtain a two-color molecule, the label transition time must occur after the first initiation and before the last coalescence.
doi:10.1371/journal.pone.0032053.g005

locations of the replication forks at the switching time (depicted by arrows). Then, if we know the labeling sequence (red followed by green in this case), we can distinguish left- from right-moving forks (forks are moving from red to green zones).

In practice, the red- and green-labeling periods are preceded by normal periods of non-fluorescent nucleotide synthesis. If each of these labeling periods is significantly longer than the duplication time of the analyzed molecules, then every molecule that is examined will show one or two types of nucleotide (but never three). All replicated molecules are collected, but only the ones that are fully labeled with fluorescent markers are kept for analysis (fully red, fully green, or red-green molecules).

The molecule-selection procedure described above–replication simulation followed by random molecule selection–can be repeated to collect a distribution of molecules. Figure 6 a shows an example of 150 red-green labeled molecules collected during a simulation of our model system (Fig. 2) using the protocol of the SMARD experiment. We simulated more molecules but kept only the ones with both labels. The red-green molecules in Fig. 6 a are organized according to their red-label content. Note that a simple visual inspection of Fig. 6 a is sufficient to obtain a general sense about the position and relative efficiency of the replication origins located in the region.

Data analysis. Figures 6 b and 6 c present three statistical "profiles" that are functions of the genome position but averaged over all the simulated molecules shown in Fig. 6 a: the local red-green ratio and the densities of replication forks in both directions. Quantities are averaged over all samples because typical experimental data sets are small (10 to 100 red-green molecules, Demczuk *et al.*, unpublished). As we shall see in the next sections, we can adapt our approach to reproduce such average quantities without having to do simulations.

Figure 6 b shows the red-green content, $\Psi(x)$, as a function of the genome position averaged over all the molecules collected in Fig. 6 a. This quantity is always between one (all red) and zero (all green) and is given by

$$\Psi(x) = \frac{1}{N_s}\sum_{i=1}^{N_s}\psi_i(x), \qquad (18)$$

where N_s is the number of samples collected and $\psi_i(x)$ is the label value (1 for red and 0 for green) of sample i at the position x. Figure 6 b clearly shows that the positions, widths and amplitudes of the red-green content function peaks correlates with the initiation zones in Fig. 2 . To a first approximation, a maximum of $\Psi(x)$ corresponds to an initiation zone, while its numerical value reflects the zone efficiency. We verified that an increase of the initiation zone width also correlates with an increase of the corresponding red-green peak width (not shown).

Another measurement that can be extracted from SMARD experiments is the position of forks along the genome. Figure 6 c shows the fork densities $\Omega_\pm(x)$ as a function of the genome position (again, the ± sign refers to right- and left-moving forks, respectively). Since the fork density is defined as number of forks per kb, it is, in the context of the SMARD experiment, given by

$$\Omega_\pm(x) = \frac{1}{N_s}\sum_{i=1}^{N_s}\omega_{i\pm}(x), \qquad (19)$$

where the local fork density $\omega_{i\pm}(x)$ is the number of forks observed in sample i in a bin of size Δx, divided by Δx. Again, the fork densities shown in Fig. 6 c were obtained from all the molecules presented in Fig. 6 a. These figures also show that the two fork densities can be used to characterize the initiation zones. For example, the position of an initiation zone approximately corresponds to the intersection of a decreasing left-moving fork density with an increasing right-moving fork density. Of course, since there are fewer forks per molecule, the fluctuations in densities are higher than the fluctuations in red-green content. Intuitively, this observation results from the fact that initiation zones are regions from which both types of forks emerge, leading to the observed positive and negative gradients of right- and left-moving fork densities across the zones. In other words, a right-moving fork is more likely to survive (not coalesce) as its moves across the zone (and vice versa for left forks). The converse situation, decreasing right-moving fork density and increasing left-density, characterizes termination zones, which are regions where coalescences are more likely to happen.

Estimating SMARD-like data from rate-equations results. Solving the rate equations (Eqs. 1 and 2) does not directly lead to quantities that we can compare to data obtained from SMARD experiments. The quantities $\Psi(x)$ and $\Omega_\pm(x)$ are not simple time averages of $f(x,t)$ and $\rho_\pm(x,t)$. In the SMARD experiment, one collects only molecules with red and green labels, which means that all of them come from DNA that was replicated during the two labeling periods. For example, that means that fragments can only be collected between ~3000 sec and ~12000 sec in the case illustrated in Fig. 5. However, the $f(x,t)$ profile obtained from our rate equations corresponds to the average of an infinite number of space-time replication events similar to the one shown in Fig. 5 but it includes information collected at all times

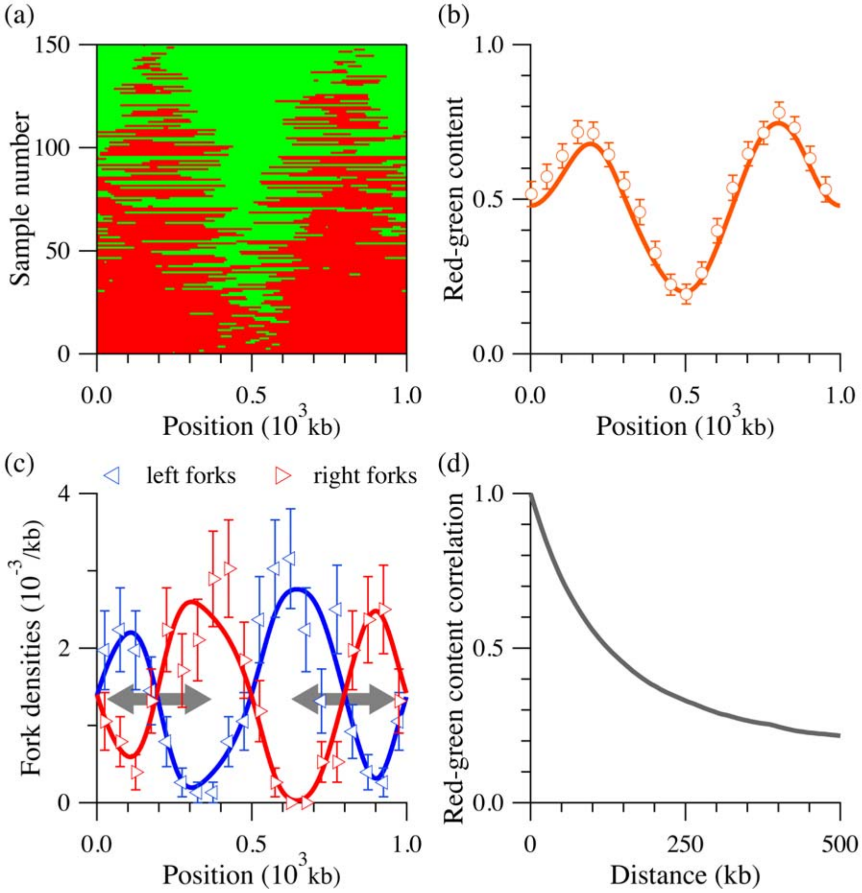


Figure 6. Simulation of SMARD experiment with comparison to rate-equation estimates. (a) Labeled molecules collected from simulations of the SMARD procedure, using the model system of Fig. 2 . Each line corresponds to a molecule as the example presented in Fig. 5 b. Molecules were organized according to their red-label content. Only molecules that were fully substituted with fluorescent nucleotides were considered for the analysis. (b) Red-green content $\Psi(x)$ of the molecules from (a) as a function of the position x along the genome (circles). A value of one (zero) means that all the molecules are red (green) labeled at a given position. The solid line was calculated using our rate equations for $f(x,t)$ (see Eq. 23). Red-green content was determined by averaging over 5 kb bins; for clarity, only one value in ten is shown. (c) Left- and right-moving fork densities $\Omega_\pm(x)$ observed in the molecules presented in (a) as a function of the position x along the genome (triangles). The fork density is defined as the number of forks per unit length at a given position (using 50 kb bins, 10 times larger than the simulation bin size). The solid line is derived from the rate equations for $\rho_\pm(x,t)$ (see Eq. 24). Gray arrows in background show the locations of initiation zones (i.e., from left to right, the intersections of increasing right-moving fork densities with decreasing left-moving fork densities). (d) Autocorrelation function of average red-green content, computed from the pool of molecules presented in (a). Since we used periodic boundary conditions, the maximum displacement is $L/2$.
doi:10.1371/journal.pone.0032053.g006

from $t=0$ to ∞. Consequently, the information prior to the first initiations and after the last coalescences that is incorporated in our rate-equation solution must be taken out to model the SMARD results. Fortunately, we can use our knowledge of the probabilities $P_s(t)$ and $P_e(t)$ to estimate $\Psi(x)$ and $\Omega_\pm(x)$.

In order to convert our calculated mean-field profile $f(x,t)$ to SMARD-like red-green content function $\Psi(x)$, we first recall that $f(x,t)$ is the average of an infinite number of single replication events similar to the one depicted in Fig. 3 a–I (f is 0 or 1 in Fig. 3 a–I, while it is a continuous number between 0 and 1 in Fig. 3 b–I and in 3 c–I). The replication fraction profile in Fig. 3 b–I is given by

$$f(x,t)=\frac{1}{N}\sum_{i=1}^{N} f_i(x,t), \tag{20}$$

where $f_i(x,t)=\{0,1\}$ is a single-event replication profile (as in Fig. 3 a–I), and N is the number of events (or simulations). The solution to the rate equations corresponds to $N\to\infty$. Equation 20 can be re-expressed as

$$\begin{aligned} f(x,t)&=\frac{1}{N}\left[\sum_{i,0<f_i(t)<1} f_i(x,t)+\sum_{i,f_i(t)=1} f_i(x,t)\right] \\ &=\frac{1}{N}\sum_{i,0<f_i(t)<1} f_i(x,t)+P_e(t), \end{aligned} \tag{21}$$

where $f_i(t)=\int_0^L f_i(x,t)dx$ is the replication fraction averaged over the whole molecule. The terms with $f_i(t)=0$ represent molecules collected at time t that have not begun to replicate. They are not included in the sum in Eq. 21 , since they each contribute 0. The terms with $f_i(t)=1$ represent molecules collected at time t that have completely replicated. Their average just gives the probability that replication has ended by that time, $P_e(t)$.

Assuming the population of cells to be perfectly asynchronous, we can collect molecules at any time t, as long as replication has started, but not ended, at time t. Consequently, our estimate of the red-green content function $\Psi(x)$ from the rate-equation solution is given by

$$\Psi(x)=\frac{\sum_{i=1}^{N_\mathrm{s}}\psi_i(x)}{N_\mathrm{s}}=\frac{\int_0^\infty \sum_{i,0<f_i(t)<1} f_i(x,t)\,dt}{N\int_0^\infty P_\mathrm{s}(t)[1-P_\mathrm{e}(t)]dt}, \tag{22}$$

where the number of samples N_s is given by the number of replication events N times the integral of the probability that DNA is actually being replicated at time t (i.e., the probability that replication has started multiplied by the probability that it has not finished). Using Eq. 21 , we can rewrite the red-green content function in a form that can be evaluated in terms of the rate-equation solution:

$$\Psi(x)=\frac{\int_0^\infty [f(x,t)-P_\mathrm{e}(t)]dt}{\int_0^\infty P_\mathrm{s}(t)[1-P_\mathrm{e}(t)]dt}. \tag{23}$$

Note that the term $P_\mathrm{e}(t)$ corrects for fully replicated molecules that are included in the calculation of $f(x,t)$ but not in $\Psi(x)$. (No correction is needed for completely unreplicated molecules since their f-value is zero.) We use Eq. 23 and the solution to the rate equations to plot the solid line in Fig. 6 b.

Similarly, the average fork density in the SMARD experiment $\Omega_\pm(x)$ is given by

$$\Omega_\pm(x)=\frac{\int_0^\infty \rho_\pm(x,t)dt}{\int_0^\infty P_\mathrm{s}(t)[1-P_\mathrm{e}(t)]dt}. \tag{24}$$

After substituting the rate-equation solution into Eq. 24 , we plot the solid lines in Fig. 6 c. In contrast with Eq. 23 , no correction for fully replicated molecules is needed in Eq. 24 since fully replicated molecules have no forks ($\rho=0$).

Figure 6 b and c compare our calculated estimates of $\Psi(x)$ and $\Omega_\pm(x)$ to simulation results. These figures demonstrate that Eqs. 23 and 24 can be used to accurately reproduce the simulated profiles obtained from experimentally typical size data set. Consequently, our model can be used to fit SMARD data in order to infer the initiation and fork velocity profiles.

One last issue that needs to be addressed is that the data points obtained from a single SMARD experiment are correlated. We can see this in Fig. 6 d, which plots $\langle\Psi(x)\Psi(x+\Delta x)\rangle$, the autocorrelation function, as a function of Δx. This means that the probability of being replicated at x is not independent of the probability of being replicated at $x\pm\Delta x$. As a consequence, the weights given each point in a fit must take into account that errors in nearby points are likely to be similar in neighboring bins.

Fitting to correlated data. Standard least-squares fitting programs assume that the statistical errors in each data point in the fit are independent. However, we have just argued that our errors show significant correlations. In order to make valid inferences about issues such as the goodness of fit, we need to take these correlations into account. To do this using standard curve-fitting routines, we linearly transform the data set to diagonalize the covariance matrix (see [45] for example). Such decorrelated data are then independent, which means that standard statistical tests (e.g., the chi-square statistic) can be used to measure the quality of a fit. Moreover, as we shall see, the diagonalization can be done in a way that evenly weights all decorrelated data (i.e., the weights can be set equal to one). Equal weights are optimal numerically for curve fitting.

Let the experimental data be expressed as a one-dimensional vector $\mathbf{d}$ that comprises the red-green profile and the fork density densities (or any other information we can extract from both the data and our rate-equation solution). The covariance matrix $\mathbf{C}$ of the data set $\mathbf{d}$ is then given by

$$\mathbf{C}(\mathbf{d})=\langle(\mathbf{d}-\langle\mathbf{d}\rangle)(\mathbf{d}-\langle\mathbf{d}\rangle)^T\rangle, \tag{25}$$

where $\langle\ldots\rangle$ represents an ensemble average over many repetitions of the experiment. The decorrelation procedure requires a matrix $\boldsymbol{\Gamma}$ that changes coordinates in the data space so that $\mathbf{C}=\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$, where the matrix $\boldsymbol{\Lambda}$ is diagonal. We say that $\boldsymbol{\Gamma}$ is a *decorrelation matrix* because the covariance matrix of the decorrelated data, denoted $\mathbf{d}^*=\boldsymbol{\Gamma}^{-1}\mathbf{d}$, is given by the diagonal matrix $\boldsymbol{\Lambda}$. Given a correlation matrix $\mathbf{C}$, many different valid decorrelation matrices can be found, as long as $\boldsymbol{\Lambda}$ is diagonal.

We can restrict the choices of decorrelation matrices by adding the constraint that all the decorrelated data points should have equal weight. This means that the diagonal matrix $\boldsymbol{\Lambda}$ can be scaled equal to the identity matrix, which implies that the decorrelation matrix $\boldsymbol{\Gamma}$ satisfies $\mathbf{C}=\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$. One way to obtain such a factorization of the correlation matrix is to perform a Cholesky decomposition of $\mathbf{C}$ such that [46]

$$\mathbf{C}=\mathbf{L}\mathbf{L}^T, \tag{26}$$

where $\mathbf{L}$ is a lower triangular matrix. The Cholesky decomposition can be performed on the correlation matrix because $\mathbf{C}$ is, by definition, symmetric and positive definite. Consequently, the Cholesky matrix $\mathbf{L}$ converts correlated data into evenly weighted decorrelated data (with all weights set to unity). Then, the following recursive procedure can be used to find the best fit of the data set:

1. Choose an initial replication scenario (initiation rate and velocity profile) that approximately reproduces the observed data $\mathbf{d}$. In order to perform a fit, the scenario must be expressed using a finite number of parameters.
2. Solve the rate equations using the current replication scenario. Estimate the data set $\widehat{\mathbf{d}}$, consisting of the red-green and fork-density profiles.
3. Perform N simulations based on the current replication scenario. Each simulation should collect the same number of fully labeled molecules as were collected during the real experiment. Analyze each simulation in the way real molecules were treated, and record the series of simulated data vectors $\mathbf{d}_i^{\mathrm{sim}}$, where the index $1<i<N$.
4. Calculate the covariance matrix of the simulated data, $\mathbf{C}(\mathbf{d}_i^{\mathrm{sim}})$. In practice, if the number of simulation runs is not large enough, the estimated covariance matrix may not be positive definite, as required to perform a Cholesky decomposition. Alternately, one can parametrize (e.g., by exponential decays) the correlations and fit any unknown parameters to simulation data. The form of the parametrized covariance matrix, denoted $\hat{\mathbf{C}}$, can chosen to ensure that $\hat{\mathbf{C}}$ is positive definite.
5. Calculate the Cholesky decomposition matrix, $\mathbf{L}$, of the parametrized covariance matrix such that $\hat{\mathbf{C}}=\mathbf{L}\mathbf{L}^T$ [46].
6. Decorrelate the observed data $\mathbf{d}$ using the Cholesky matrix. The decorrelated data, denoted $\mathbf{d}^*$, are given by $\mathbf{d}^*=\mathbf{L}^{-1}\mathbf{d}$.
7. Fit the decorrelated data $\mathbf{d}^*$ with the decorrelated solution of our rate-equations, $\widehat{\mathbf{d}}^*=\mathbf{L}^{-1}\widehat{\mathbf{d}}$. The fit searches for the replication scenario that minimizes the difference between

the decorrelated data vectors $\mathbf{d}^*$ and $\widehat{\mathbf{d}}^*$ (where the weights of all data sets components are equal and set to unity). The correlated fit solution is given by $\widehat{\mathbf{d}} = \mathbf{L}\widehat{\mathbf{d}}^*$.

8. Repeat, starting from Step 2, using the latest fit result as the current replication scenario, until the solution converges.

Fit example. We now apply the correlated data fitting procedure described above to a real SMARD data set. The data we use here and all the experimental details related to their collection can be found in Demczuk *et al.*, unpublished. In this paper, the SMARD technique was used to study DNA replication in mouse bone marrow pro-B cells at different developmental stages. The study was performed on four adjacent restriction fragments that cover about ~ 1.4 Mb of the genome. Because the fragments come from a much longer genome, we did not use periodic boundary conditions but instead modeled explicitly the injection of outside forks into the studied region.

In Fig. 7 , we present global fits to six different fragments (from Demczuk *et al.*, unpublished). The term "global" here means that all the fragments are simultaneously fit by a common, or global, set of parameters. Fragments 1 to 4 cover the studied region in unrearranged normal pro-B cells (left side of Fig. 7). The last two fragments (3′ and 4′) come from a clonal population of cells containing a genomic rearrangements within fragment 3 (right side of Fig. 7). The rearrangement of fragment 3 into 3′ consist in a genomic deletion of approximatively 65 kb (located at 68 kb from the right end of fragment 3, see dashed lines in Fig. 7).

In fitting the experimental data, we made the following assumptions about the replication scenario:

1. Based on the normal cell red-green content profile (left side of Fig. 7 a), we assumed that two initiations zones are present (around 250 kb and 1150 kb). Each zone has three parameters that describe the position, width, and initiation rate of the zone. Another parameter defines a constant background of initiation (this parameter was added because low levels of initiations were observed outside the initiation regions). Finally, two other parameters describe fork injection rates at the boundaries of the modeled region (see filled symbols in Fig. 7 d).
2. For practical reasons, we assumed that the shape for the initiation zones was a rounded box, such as the ones shown in Fig. 7 d. As we see in Fig. 7 , the red-green content profile is not too sensitive to the precise shape of the initiation zones (e.g., the red-green content maxima have smoother edges than their corresponding boxy initiation zones).
3. We also assume that the initiation profile does not change with time during the S phase. Time-dependent profiles were considered but did not affect significantly the fit (unpublished observation).
4. Data sets from unrearranged and rearranged alleles were assumed to have the same initiation rates except within fragments 3/3′. The linear red-green content profiles and the corresponding fork densities of fragments 3′ and 4′ indicate that these fragments are almost always replicated by left-moving forks coming from the right side of fragment 4′. We thus assumed that the initiation profile of the deleted allele is the same as the one of undeleted allele except for the absence of the second initiation zone located within the deleted region (compare fragments 3 and 3′ in Fig. 7 d).
5. We assumed a constant velocity throughout the four fragments. However, the experimental results presented in Demczuk *et al.*, unpublished, indicate that forks propagated at different speeds in these two experiments (probably caused by differences in the growing rate of the cultured cells in the two experiments). Therefore, we used two fork speed parameters, one for fragments 1 to 4 and another one for fragments 3′ and 4′.

The hypothetical replication described above comprises 11 free parameters that can be adjusted throughout a fitting routine (6 for the two initiation zones, 1 for the background initiation rate, 2 for forks coming from outside the modeled region, and 2 for velocities in both cell types). Using that hypothesis, we followed the fitting procedure described in Section to perform a global fit of the SMARD data collected from the six fragments. The fitted $\Psi(x)$ and $\Omega(x)$ profiles are shown as solid lines in Fig. 7 a–c. The best-fit results are illustrated in Fig. 7 d as an initiation-rate curve. Note that our rate-equation system has to be solved two times for a given set of parameters (with and without the second initiation zone for normal and clone cells, respectively).

Since determining the replication program was the aim of the experiment, the quality of the fit cannot be directly compared to the "actual" replication program. However, SMARD provides information that was not used for the fit. Hence, it is possible to verify that the result of the fit are consistent with this additional information. First, the fitted fork velocities we obtained are 0.045 kb/sec and 0.023 kb/sec (both ± 0.003 kb/sec) for the normal and clonal data set. The corresponding experimental values are 0.041 kb/sec and 0.024 kb/sec (Demczuk *et al.*, unpublished). Considering the small sample sizes used to obtain these fork velocities (from 11 to 57 fully labeled molecules only, depending on the fragment, Demczuk *et al.*, unpublished), we evaluated from simulations the statistical errors for the measured fork velocities ($\approx \pm 10\%$). (Experimentally, the fork velocity within a fragment is calculated as $v = \ell / \left(n_{\rm f} t_{\rm rep}\right)$, where ℓ is the fragment length, $n_{\rm f}$ the average number of forks observed per fragments, and $t_{\rm rep}$ the replication time of the fragment. The replication time is given by $t_{\rm rep} = t_{\rm pulse} n_{\{r,g\}} / n_{\{r,g\}} + n_{rg}$, where $n_{\{r,g\}}$ is the number of fully red (or green) labeled fragments while $n_{\rm rg}$ is the number of fully labeled fragments that have incorporated both labels (Demczuk et al., unpublished).) Thus, our fitted values nicely agree with the experiments. Second, the position of the second initiation zone, [1.11 Mb, 1.17 Mb] (± 0.01 Mb), is almost completely located within the genomic deletion region of fragment 3, which is found between [1.12 Mb, 1.18 Mb]. (Remember that we did not use the deletion location to restrict the second initiation zone position while fitting.)

Our fit result has a reduced chi-square statistic of $\chi^2 = 1.13 \pm 0.05$ with 694 degrees of freedom. This high χ^2 value is due to the simplistic initiation function we used. For example, a more complicated initiation function could be used to obtain a better fit of the red-green content profiles (e.g., we could use a higher initiation rate at the right side of fragment 2 or a different shape for the zone in fragment 3). Nevertheless, we believe that the simple replication scenario used here captures the most important features of the data set. Moreover, when we use the fit result to perform simulations of the SMARD experiment, we obtain statistics about the initiation/coalescence events and the replication time of each fragments that agree with the experimental values (Demczuk *et al.*, unpublished).

Discussion

Over the years, various experimental approaches have been used to measure the absolute and relative efficiencies of origin firing in eukaryotic cells. However, the efficiency of origin firing does not encapsulate all the information required to understand how DNA origins of replication are regulated. Since eukaryotic genomes contain large numbers of origins, understanding their regulation requires a quantitative analysis of the dynamics of

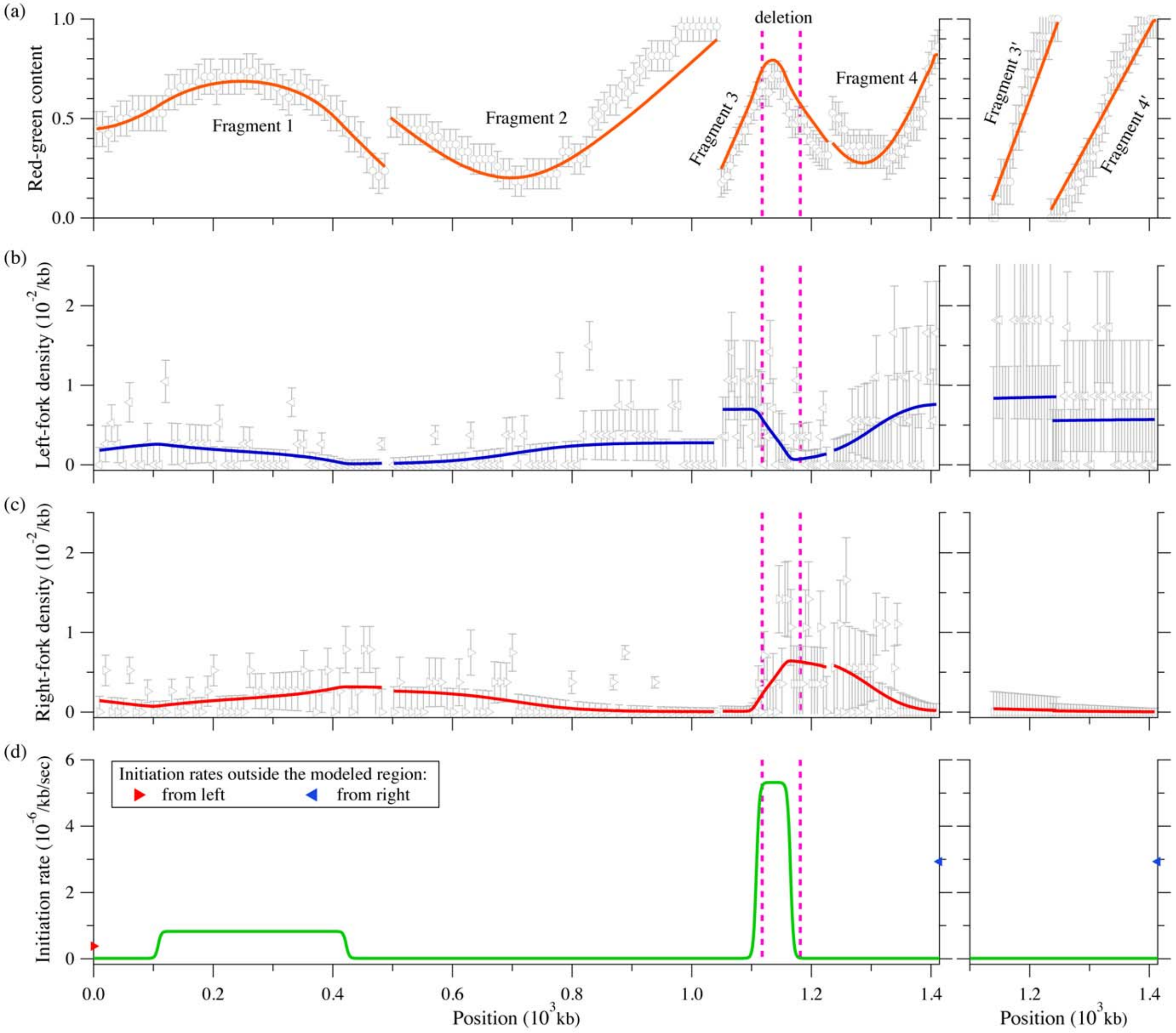


Figure 7. SMARD analysis of DNA replication in mouse bone marrow pro-B cells. The left side presents the data collected from four fragments covering a ~ 1.4 Mb region in normal cells. The right side shows data obtained from clone cells where the genome sequence was rearranged (65 kb was deleted from the genome). ÊThe deletion is located between the two dashed lines on the left side graphs. Only the equivalent of fragments 3 and 4 from normal cells was studied in the clonal population. Symbols represent experimental data while solid lines refer to the solution of our rate-equation system. (a) Red-green content $\Psi(x)$ obtained from Eqs. 18 (symbols) and 23 (solid lines). (b, c) Left- and right-moving fork densities $\Omega_{\pm}(x)$ given by Eqs. 19 (symbols) and 24 (solid lines). (d) Best fit result for the initiation rate $I(x)$ (solid lines) and boundary fork injection rates (symbols) used to solve our rate-equations. The best-fit fork velocities we obtained were 0.045 kb/sec and 0.023 kb/sec for normal and clonal cell populations, respectively. Errors bars in (a, b, c) were obtained from simulations of the best-fit replication scenario.
doi:10.1371/journal.pone.0032053.g007

origin firing along the genome and across S phase. Achieving this goal requires comprehensive data sets about DNA replication across large genomic regions, as well as mathematical procedures for the analysis of complex data sets.

In this manuscript, we present a new set of rate equations that can be used to calculate the firing rate of DNA origin of replication using multiple sets of data (temporal order of replication, fork density, replication time). Our mathematical procedure is versatile and allows the analysis of complex data sets obtained using various experimental approaches (SMARD, microarrays, etc.). This is possible because our model follows the spatial and temporal evolution of several replication factors. In contrast, previous procedures have mostly relied on the analysis of individual parameters of DNA replication that can be modeled with limited detail (e.g., timing of replication). The main advantage of this technique is that the rate-equation solution corresponds to the exact mean-field replication program. Our approach thus provides more precise information about average replication kinetics than Monte Carlo simulations. It is faster, too. As discussed previously, simulation remains the appropriate technique for estimating statistical fluctuations of replication-related quantities. Since average replication kinetics is often the only information obtainable from experiments, our model is, in many practical cases, sufficient to reproduce experimental data. For these reasons, our mathematical procedure makes it possible to perform a faster, and more thorough, analysis of the process of DNA replication initiation and of its regulation in complex eukaryotes.

Although our procedure can be used to analyze data sets obtained with different experimental approaches, we validated it using results of recent SMARD experiments performed across a 1.4 Mb region which spans the mouse immunoglobulin heavy chain locus (Demczuk *et al.*, unpublished). We chose these experiments because, besides providing the data sets used in all the calculations, SMARD provided us with additional information that could be directly compared with the predictions of the procedure (e.g., the location of initiation events and fork collisions, the number of molecules containing such events, and the average number of events per molecules). The close match between calculated and experimental data sets indicates that our procedure can be used to make valuable inferences about various aspects of DNA replication in eukaryotes, with the calculations taking only modest computer resources. The usefulness of our model was illustrated by the series of fits of SMARD data we performed in Demczuk *et al.*, unpublished.

In Demczuk *et al.*, unpublished, the methods presented here implied that origin firing within the mouse Igh locus is compatible with the stochastic firing of origins throughout S phase, with a rate that varies along the locus. The Igh locus is divided into domains of similar firing rates, and the rate of firing within these domains is developmentally regulated. These observations contrast notably with results obtained in budding yeast, where the rate of firing varies from origin to origin and coordination in origin activity has not been observed [18]. Moreover, this approach allowed us to study various aspects of the developmental regulation of origin activity during B cell development.

In summary, the mathematical procedure described in this study has already provided new insights on the regulation of DNA replication initiation in mammalian cells and makes possible the study of additional phenomena such as replication time in the presence of fork velocities that depend on genome location or the impact of a correlation between initiation rates and fork density. Our method is thus a natural starting point for investigating checkpoint mechanisms where, for example, the cell regulates the local or global replication activity in response to various intra- or extracellular feedback signals.

Supporting Information

Figure S1 Space-time diagram of replication with inhomogeneous fork speeds. The space-time point (X,T) is replicated by an initiation that occurred within the shaded area (e.g., initiation A). By contrast, initiation B will replicate the location X but only at a time $t>T$. The inset defines symbols that refer to different portions of the shaded area. Note that $\Delta=\Delta_-+\Delta_+$.
(TIF)

Supporting Information S1 Ending probability (homogeneous case). Modeling fork injection at boundaries.
(PDF)

Acknowledgments

We thank N. Rhind and S. Jun for their careful reading of our manuscript.

Author Contributions

Conceived and designed the experiments: PN. Performed the experiments: PN. Analyzed the data: MGG JB. Wrote the paper: MGG JB.

References

1. Kaufmann WK (2010) The human intra-S checkpoint response to UVC-induced DNA damage. Carcinogenesis 31: 751–65.
2. Branzei D, Foiani M (2005) The DNA damage response during DNA replication. Curr Opin Cell Biol 17: 568–75.
3. Jun S, Zhang H, Bechhoefer J (2005) Nucleation and growth in one dimension. I. The generalized Kolmogorov-Johnson-Mehl-Avrami model. Phys Rev E 71: 011908.
4. Jun S, Bechhoefer J (2005) Nucleation and growth in one dimension. II. Application to DNA replication kinetics. Phys Rev E 71: 011909.
5. Norio P, Kosiyatrakul S, Yang Q, Guan Z, Brown NM, et al. (2005) Progressive activation of DNA replication initiation in large domains of the immunoglobulin heavy chain locus during B cell development. Mol Cell 20: 575–587.
6. Norio P, Schildkraut CL (2004) Plasticity of DNA replication initiation in Epstein-Barr virus episomes. PLoS Biol 2: 816–833.
7. Norio P, Schildkraut CL (2001) Visualization of DNA replication on individual Epstein-Barr virus episomes. Science 294: 2361–2364.
8. Katsuno Y, Suzuki A, Sugimura K, Okumura K, Zineldeen DH, et al. (2009) Cyclin A-Cdk1 regulates the origin firing program in mammalian cells. P Natl Acad Sci USA 106: 3184–9.
9. Courbet S, Gay S, Arnoult N, Wronka G, Anglana M, et al. (2008) Replication fork movement sets chromatin loop size and origin choice in mammalian cells. Nature 455: 557–60.
10. Huvet M, Nicolay S, Touchon M, Audit B, d'Aubenton Carafa Y, et al. (2007) Human gene organization driven by the coordination of replication and transcription. Genome Research 17: 1278–85.
11. Herrick J, Bensimon A (2008) Global regulation of genome duplication in eukaryotes: an overview from the epiuorescence microscope. Chromosoma 117: 243–60.
12. Lebofsky R, Heilig R, Sonnleitner M, Weissenbach J, Bensimon A (2006) DNA replication origin interference increases the spacing between initiation events in human cells. Mol Biol Cell 17: 5337–45.
13. Nieduszynski CA, Blow JJ, Donaldson AD (2005) The requirement of yeast replication origins for pre-replication complex proteins is modulated by transcription. Nucleic Acids Res 33: 2410–20.
14. Alexandrow MG, Hamlin JL (2005) Chromatin decondensation in S-phase involves recruitment of Cdk2 by Cdc45 and histone H1 phosphorylation. J Cell Biol 168: 875–86.
15. Marheineke K, Hyrien O (2001) Aphidicolin triggers a block to replication origin firing in Xenopus egg extracts. J Biol Chem 276: 17092–100.
16. Marheineke K, Hyrien O (2004) Control of replication origin density and firing time in Xenopus egg extracts: role of a caffeine-sensitive, ATR-dependent checkpoint. J Biol Chem 279: 28071–81.
17. Patel PK, Arcangioli B, Baker SP, Bensimon A, Rhind N (2006) DNA replication origins fire stochastically in fission yeast. Mol Biol Cell 17: 308–16.
18. Yang SCH, Rhind N, Bechhoefer J (2010) Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. Molecular Systems Biology 6: 404.
19. Gauthier MG, Herrick J, Bechhoefer J (2010) Defects and DNA replication. Phys Rev Lett 104: 218104.
20. Gauthier MG, Bechhoefer J (2009) Control of DNA replication by anomalous reaction-diffusion kinetics. Phys Rev Lett 102: 158104.
21. Yang SCH, Gauthier MG, Bechhoefer J Computational methods to study kinetics of DNA replication, in DNA Replication: Methods and Protocols, Humana Press, chapter 32. pp 555–574.
22. Bechhoefer J, Marshall B (2007) How Xenopus laevis replicates DNA reliably even though its origins of replication are located and initiated stochastically. Phys Rev Lett 98: 098105.
23. Zhang H, Bechhoefer J (2006) Reconstructing DNA replication kinetics from small DNA fragments. Phys Rev E 73: 051903.
24. Yang SCH, Bechhoefer J (2008) How Xenopus laevis embryos replicate reliably: Investigating the random-completion problem. Phys Rev E 78: 041917.
25. Herrick J, Jun S, Bechhoefer J, Bensimon A (2002) Kinetic model of DNA replication in eukaryotic organisms. J Mol Biol 320: 741–750.
26. Kolmogorov A (1937) A statistical theory for the recrystallization of metals. Bull Acad Sc USSR, Phys Ser 1 1: 335.
27. JohnsonWA, Mehl FL (1939) Reaction kinetics in processes of nucleation and growth. Trans AIME 135: 416.
28. Avrami M (1939) Kinetics of phase change. I General theory. J Chem Phys 7: 1103.
29. Avrami M (1940) Kinetics of phase change. II Transformation-time relations for random distribution of nuclei. J Chem Phys 8: 212.
30. Avrami M (1941) Granulation, phase change, and microstructure - Kinetics of phase change. III. J Chem Phys 9: 177.

31. Blow JJ, Ge XQ (2009) A model for DNA replication showing how dormant origins safeguard against replication fork failure. EMBO Rep 10: 406–412.
32. Lygeros J, Koutroumpas K, Dimopoulos S, Legouras I, Kouretas P, et al. (2008) Stochastic hybrid modeling of DNA replication across a complete genome. P Natl Acad Sci USA 105: 12295–300.
33. Goldar A, Labit H, Marheineke K, Hyrien O (2008) A Dynamic stochastic model for DNA replication Initiation in early embryos. PLoS ONE 3: e2919.
34. Spiesser TW, Klipp E, Barberis M (2009) A model for the spatiotemporal organization of DNA replication in Saccharomyces cerevisiae. Mol Genet Genomics 282: 25–35.
35. de Moura APS, Retkute R, Hawkins M, Nieduszynski CA (2010) Mathematical modelling of whole chromosome replication. Nucleic Acids Res 38: 5623–33.
36. Bensimon A, Simon A, Chiffaudel A, Croquette V, Heslot F, et al. (1994) Alignment and sensitive detection of DNA by a moving interface. Science 265: 2096–8.
37. Herrick J, Bensimon A (1999) Imaging of single DNA molecule: applications to high-resolution genomic studies. Chromosome Res 7: 409–23.
38. Raghuraman MK, Winzeler EA, Collingwood D, Hunt S, Wodicka L, et al. (2001) Replication dynamics of the yeast genome. Science 294: 115–21.
39. Feng W, Collingwood D, Boeck ME, Fox LA, Alvino GM, et al. (2006) Genomic mapping of singlestranded DNA in hydroxyurea-challenged yeasts identifies origins of replication. Nat Cell Biol 8: 148–55.
40. Heichinger C, Penkett CJ, Bähler J, Nurse P (2006) Genome-wide characterization of fission yeast DNA replication origins. EMBO J 25: 5171–9.
41. Conti C, Saccà B, Herrick J, Lalou C, Pommier Y, et al. (2007) Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. Mol Biol Cell 18: 3059–3067.
42. Frank FC (1974) Nucleation-controlled growth on a one-dimensional growth of finite length. J Cryst Growth 22: 233–236.
43. Vilenchik MM, Knudson AG (2003) Endogenous DNA double-strand breaks: production, fidelity of repair, and induction of cancer. P Natl Acad Sci USA 100: 12871–6.
44. Woodward A, Göhler T, Luciani M, Oehlmann M, Ge X, et al. (2006) Excess Mcm2–7 license dormant origins of replication that can be used under conditions of replicative stress. J Cell Biol 173: 673.
45. Tellinghuisen J (1994) On the least-squares fitting of correlated data: Removing the correlation. Journal of Molecular Spectroscopy 165: 255–264.
46. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) Numerical Recipes: The Art of Scientific Computing, Third Edition. New York: Cambridge University Press.

Supporting Information for: Modeling inhomogeneous DNA replication kinetics

Michel G. Gauthier [1], Paolo Norio[2,3], John Bechhoefer [1,*]

1 Department of Physics, Simon Fraser University, Burnaby, British Columbia, Canada, V5A 1S6

2 Department of Oncology, Montefiore Medical Center, Moses Division, Bronx, NY, USA

3 Department of Medicine, Albert Einstein College of Medicine, Bronx, NY, USA

$\ast$ E-mail: johnb@sfu.ca

Ending probability (homogeneous case)

To test our approximation of the replication ending probability $\widetilde{P}_\mathrm{e}(t)$ presented in Eq. 12, we consider here a test case where $P_\mathrm{e}(t)$ can be solved exactly—the case of homogeneous $I(x,t) = I$ and $v(x,t) = v$. For this case, we previously demonstrated that the replication fraction is given by [1]

$$f(t) = 1 - e^{-Ivt^2}\,, \tag{S1}$$

and the fork densities by

$$\rho_\pm(t) = \rho(t) = \frac{1}{2v}\frac{\partial f(t)}{\partial t} = Ite^{-Ivt^2}\,, \tag{S2}$$

where the genome length L is assumed to be very large compared to $\sqrt{v/I}$. From Eq. 5, the coalescence time probability density function is given by

$$\phi_\mathrm{c}(t) = \frac{2v\rho(t)^2}{N_\mathrm{c}(1-f(t))} = \frac{4}{\sqrt{\pi}}\,(Iv)^{3/2}\,t^2e^{-Ivt^2}\,, \tag{S3}$$

where the number of coalescences per cell cycle is given by $N_\mathrm{c} = \int_0^\infty\int_0^L \frac{2v\rho^2}{(1-f)}\,dx\,dt = L\sqrt{I\pi/v}/2$. The cumulative density function of the coalescence times, $\Phi_\mathrm{c}(t)$, is then

$$\begin{aligned}\Phi_\mathrm{c}(t) &= \int_0^t \phi_\mathrm{c}(t')dt' = \mathrm{erf}(\sqrt{Iv}t) - 2t\sqrt{\frac{Iv}{\pi}}e^{-Ivt^2} \\ &= \mathrm{erf}(t^*) - \frac{2t^*}{\sqrt{\pi}}e^{-t^{*2}} = \Phi_\mathrm{c}(t^*)\,,\end{aligned} \tag{S4}$$

where $t^* = \sqrt{Iv}t$ is a dimensionless time parameter. The probability that replication has finished by time t^* can then be calculated as the probability that N_c coalescences have occurred by time t^*,

$$P_\mathrm{e}(t^*) = [\Phi_\mathrm{c}(t^*)]^{N_\mathrm{c}}\,. \tag{S5}$$

If N_c is large, we can re-write the exact solution as

$$\begin{aligned}P_\mathrm{e}(t^*) &= \exp\{-N_\mathrm{c}[1-\Phi_\mathrm{c}(t^*)]\} \qquad \text{(S6)}\\ &= \exp\left\{\frac{-2N_\mathrm{c}t^*e^{-t^{*2}}}{\sqrt{\pi}}\left[1 + \frac{\sqrt{\pi}\,\mathrm{erfc}(t^*)}{2t^*e^{-t^{*2}}}\right]\right\}.\end{aligned}$$

This exact solution may be compared to our approximation in Eq. 12,

$$\widetilde{P}_\mathrm{e}(t^*) = \exp\left\{\frac{-2N_\mathrm{c}t^*e^{-t^{*2}}}{\sqrt{\pi}(1-e^{-t^{*2}})}\right\}. \tag{S7}$$

Both curves are sigmoidal functions interpolating between 0 and 1, as one expects. A numerical comparison of Eqs. S6 and S7 (not shown) shows that the maximum difference between the two expressions decreases as N_c increases.

Modeling fork injection at boundaries

Here, we calculate the density of forks at the boundary of the region under study. The forks originate from the region outside the given boundary, and we thus do not have direct information about their origins. Figure S1 presents a space-time diagram of the replication situation assuming inhomogeneous initiation rates and the fork velocities. In this situation, the probability that an arbitrary point X along the genome remains unreplicated at time T equals the probability that no initiation occurred in the shaded space-time region shown in Fig. S1. That region is the set of space-time locations at which an initiation would passively replicate the point (X, T) in the diagram. Thus, the replication fraction is given by

$$f(X,T) = 1 - \exp\left\{ - \iint_{x,t\in\triangle} I(x,t)\, dx\, dt \right\} , \tag{S8}$$

where $\triangle$ represents the shaded area in Fig. S1 (the shaded region has curved boundaries because of the inhomogeneity of the velocity profile). Using the point X to define a boundary (dashed line) between two regions Δ_- and Δ_+, the rate of replication at the boundary is the sum of two contributions

$$\begin{aligned} \frac{\partial f(X,T)}{\partial t} &= [1 - f(X,T)] \int_{x\in\triangle} I(x,t)\, dx\, , \\ &= [1 - f(X,T)] \int_{x\in\triangle_-} I(x,t)\, dx \\ &\quad + [1 - f(X,T)] \int_{x\in\triangle_+} I(x,t)\, dx\, , \end{aligned} \tag{S9}$$

where $\triangle_\pm$ represents the right (+) and left (−) portions of the shaded area in Fig. S1 (see legend). By analogy with Eq. 1, the fork densities at X and T are given by

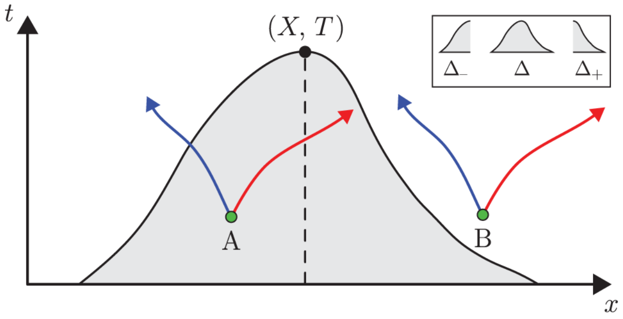
$$\rho_\pm(X,T) = \frac{[1 - f(X,T)]}{v_\pm(X,T)} \int_{x\in\triangle_\mp} I(x,t)\, dx\, , \tag{S10}$$

where $\rho_\pm$ and $\triangle_\mp$ have opposite indices because right-moving forks at $x = X$ come from the left side of the shaded area and vice versa. Equations 14 and 15 are a special case (constant I and v) of Eq. S10.

References

1. Jun S, Zhang H, Bechhoefer J (2005) Nucleation and growth in one dimension. I. The generalized Kolmogorov-Johnson-Mehl-Avrami model. Phys Rev E 71: 011908.

Supplementary Figure S1. Space-time diagram of replication with inhomogeneous fork speeds. The space-time point (X, T) is replicated by an initiation that occurred within the shaded area (e.g., initiation A). By contrast, initiation B will replicate the location X but only at a time $t > T$. The inset defines symbols that refer to different portions of the shaded area. Note that $\Delta = \Delta_- + \Delta_+$.


t
(X, T)
Δ₋
Δ
Δ₊
A
B
x