Genome replication in asynchronously growing microbial populations

Florian Pflug[1], Deepak Bhat[1,2], Simone Pigolotti[1*]
[1]*Biological Complexity Unit,*
Okinawa Institute of Science and Technology Graduate University,
Onna, Okinawa 904-0495, Japan and
[2]*Department of Physics, School of Advanced Sciences,*
Vellore Institute of Technology, Vellore, Tamil Nadu, India

Biological cells replicate their genomes in a well-planned manner. The DNA replication program of an organism determines the timing at which different genomic regions are replicated, with fundamental consequences for cell homeostasis and genome stability. Qualitatively, in a growing cell culture, one expects that genomic regions that are replicated early should be more abundant than regions that are replicated late. This abundance pattern can be experimentally measured using deep sequencing. However, a general quantitative theory to explain these data is still lacking. In this paper, we predict the abundance of DNA fragments in asynchronously growing cultures from any given stochastic model of the DNA replication program. As key examples, we present stochastic models of the DNA replication programs in *Escherichia coli* and in budding yeast. In both cases, our approach leads to analytical predictions that are in excellent agreement with experimental data and permit to infer key information about the replication program. In particular, our method is able to infer the locations of known replication origins in budding yeast with high accuracy. These examples demonstrate that our method can provide insight into a broad range of organisms, from bacteria to eukaryotes.

I. INTRODUCTION

The genome of an organism contains precious information about its functioning. Genomes need to be reliably and quickly replicated for cells to pass biological information to the next generation. Replication of a genome proceeds according to a certain plan, termed the "replication program" [1–4]. For example, most bacteria have a circular genome, where two replisomes initiate replication by binding at the same origin site [5]. They replicate the genome in opposite directions. Replication is completed when they meet, after each of them has copied approximately half of the genome. The replication program is carefully orchestrated, but not completely deterministic. Stochasticity is particularly relevant in eukaryotes and archaea, where many origin sites are present in each chromosome [6, 7]. Replication can initiate at these origin sites at different times. These times are characterized by some degree of randomness [8, 9] and, depending on the circumstances, some origin sites may not be activated at all [1].

Replication programs must be coordinated with the cell cycle in some way. For example, in bacteria, the moment in the cell cycle in which replication is initiated is carefully timed [10–12]. In eukaryotes, replication takes place at a well-defined stage of the cell cycle (the S phase), during which different genomic regions are replicated at different times in a controlled manner [8]. The interplay between the replication program and the cell cycle is crucial when trying to infer the replication program from experimental observations. Many experiments study samples in which all cells are approximately at the same stage of the cell cycle. This can be achieved by either arresting the cell cycle at a certain stage or by cell sorting [4]. The fraction of these synchronized cells that have copied each genome location can be then measured using deep sequencing. This approach has been extensively used to study the eukaryotic replication program [4, 13, 14].

An alternative method is to measure the abundance of DNA fragments in asynchronous, exponentially growing populations. This approach, traditionally called marker frequency analysis [15], has been extensively applied to bacterial replication, for example to study *Escherichia coli* mutant lacking genes that assist DNA replication [16–18] and artificially engineered *E.coli* strains with multiple replication origins [19]. The asynchronous approach is experimentally much simpler and avoids potential artifacts caused by the cell cycle arrest or cell sorting [4]. Progress in DNA sequencing have made these experiments high-throughput and relatively inexpensive. However, the DNA sampled in these experiments originates from a mixture of cells at different stages in their life cycle, rendering the theoretical interpretation of such data problematic [20].

Theoretical approaches have attempted to describe measurements in asynchronous populations. However, these approaches either neglect stochasticity, or are limited to specific model systems. For example, we recently proposed a stochastic model describing DNA replication in growing *E.coli* populations [21]. A broader range of bacterial systems have been studied assuming a deterministic replication program [22, 23]. Finally, a model of the replication program in budding yeast adopted the working hypothesis that cells in an asynchronous population are at random, uniformly distributed stages in their cell cycle [3].

* simone.pigolotti@oist.jp

In this paper, we develop a general theory to infer the replication program from sequencing of an asynchronously growing population of cells. Our theory builds upon classic results on age-structured populations [15, 24–27], that we extend to populations of stochastically replicating genomes. Our approach requires minimal assumptions on the replication dynamics. In particular, it allows for a stochastic replication program and equally applies to bacteria and eukaryotes. We apply our method to experimental data from *E. coli* and budding yeast. In the case of *E. coli*, we propose and exactly solve a model of DNA replication that incorporates stochastic stalling. The fit of the solution to experimental data sheds light on recently observed oscillations of bacterial replisome speed. In the case of yeast, our approach permits to reliably infer the location of replication origins.

The paper is organized as follows. In Section II, we present our general theory. In Section III, we apply our framework to DNA replication in *E. coli*. In Section IV, we consider eukaryotic DNA replication, with budding yeast as an example. In Section V, we present possible extensions of our theory. Section VI is devoted to conclusions and future perspectives.

II. DNA REPLICATION IN ASYNCHRONOUSLY GROWING POPULATIONS

A. General theory

We consider a large, growing population of cells. We call $N_c(t)$ the number of cells that are present in the population at time t. Each cell may contain multiple genomes, some complete and other undergoing synthesis (incomplete). We denote by $N_g(t)$ the the total number of complete and incomplete genomes in the population.

Our theory is based on the following assumptions. The number of cells grows exponentially: $N_c(t) \propto \exp(\Lambda t)$, where Λ is the exponential growth rate. The population grows in a steady, asynchronous manner. This assumption implies that the average number of genomes per cell must remain constant, and therefore $N_g(t) \propto \exp(\Lambda t)$. Genomes in the population are immortal, i.e. we neglect the rate at which they might be degraded. All genomes in the population are statistically identical, in the sense that they are all characterized by the same stochastic replication program. These assumptions are realistic in common experimental situations. Moreover, as we shall discuss in Section V, some of them can be readily relaxed, if necessary.

We now assign an age to each genome in the population. To this aim, we conventionally set the birth time of a daughter genome at the start of the replication process that generates it from a parent genome, see Fig. 1a. We note that the distinction between a "parent" and a "daughter" genome is somewhat arbitrary, as each of them is made up of a preexisting strand and a newly copied complementary one. Since genomes are immortal, the probability density of new genomes is proportional to $\dot{N}_g(t)$, which is proportional to $\exp(\Lambda t)$ as well. It follows that the distribution of ages τ of genomes in the population must be proportional to $\dot{N}_g(t-\tau)$. From this fact, we conclude that the distribution of genome ages in the population is

$$P(\tau) = \Lambda e^{-\Lambda \tau}. \tag{1}$$

We now look into the DNA replication program in more details. In bacteria, replication is carried out by a pair of replisomes, that bind at a specific genome site (the origin) and proceed in opposite directions, each replicating both strands. DNA replication is substantially more complex in eukaryotes, where a large number of replisomes replicate the same chromosome, and initiation site might be activated stochastically. We summarize the effect of all these complex processes into the probability $f(x,\tau)$ that the genome location x has already been copied in a genome of age τ. By definition, $f(x,\tau)$ is a non-decreasing function of τ. Our assumption that genomes are statistically identical means that all genomes are characterized by the same $f(x,\tau)$. Examples of deterministic and stochastic replication programs are represented in Fig. 1b and 1c, respectively.

We now define the probability $\mathcal{P}(x)$ that a randomly chosen genome in the population contains the genome location x. We remark that $\mathcal{P}(x)$ is not necessarily normalized to one when integrated over the entire genome. Its normalized counterpart represents the probability density that a randomly chosen genome fragment in the population originates from genomic location x. For this reason, we call $\mathcal{P}(x)$ the DNA abundance distribution. The DNA abundance distribution can be experimentally measured using deep sequencing.

By using Eq. (1) and integrating over all genome ages, we find that $\mathcal{P}(x)$ is related with $f(x,\tau)$ by

$$\mathcal{P}(x) = \int_0^\infty d\tau\, \Lambda e^{-\Lambda\tau} f(x,\tau), \tag{2}$$

see Fig. 1d. To find a more transparent expression, we introduce the probability density $\psi(x,\tau) = \partial_\tau f(x,\tau)$. Substituting this definition into Eq. (2) and integrating by parts we obtain

$$\mathcal{P}(x) = \left\langle e^{-\Lambda\tau} \right\rangle_x , \tag{3}$$

where $\langle \dots \rangle_x = \int_0^\infty d\tau \dots \psi(x,\tau)$ is the average over the distribution of replication ages. Equation (2), or equivalently Eq. (3), is the basis of our approach.

In our derivation, we intentionally avoided modeling the specific dynamics of the cell cycle, how it is regulated, and how it is coordinated with the DNA replication program. Our theory is rigorously valid independently of these aspects, provided that our initial assumptions hold.

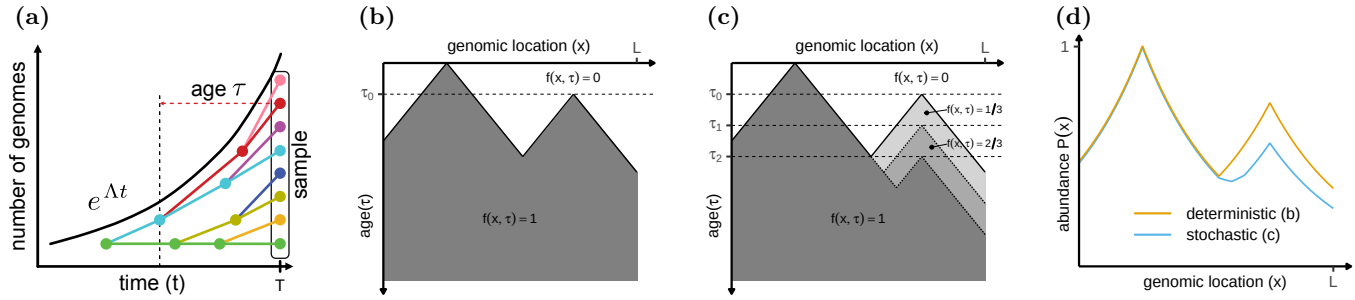

FIG. 1. DNA abundance in an asynchronous, exponentially growing microbial population. **(a)**. Total number of genomes (black line) and genealogy of individual genomes (colored tree). Nodes in the tree represent replication initiations. Such events leave the template unchanged and create a new genome (differently colored descendant) with initial age $\tau = 0$. **(b)**. Example of a deterministic replication program $f(x, \tau)$ on a linear genome in which replication is initiated at two origins, one firing at age $\tau = 0$ and one at age $\tau = \tau_0$, and proceeds deterministically with constant speed. **(c)**. A stochastic version of the replication program in (b), in which replication at the second origin is initiated randomly at either τ_0, τ_1 or τ_2. **(d)**. DNA abundance distribution arising from replication programs (b) and (c) predicted by Eq. (2).

B. Deterministic limit

In the simple case where replication proceeds deterministically and the replisome speed is a function of its position on the DNA, one has

$$\psi(x, \tau) = \delta\left(\tau - \tau_0 - \int_{x_0}^{x} \frac{dx'}{v(x')}\right) \tag{4}$$

where $v(x)$ is the replisome speed at position x, x_0 is the coordinate of the replication origin for the replisome that copied position x, and τ_0 is the firing age for that origin. The speed v might take positive or negative values depending on whether the replisome proceeds in the positive genome direction. It follows from Eqs. (3) and (4) that

$$\mathcal{P}(x) = \exp\left(-\Lambda\tau_0 - \Lambda \int_{x_0}^{x} \frac{dx'}{v(x')}\right). \tag{5}$$

Solving for the speed, we obtain

$$v(x) = -\Lambda\left[\frac{d}{dx}\ln\mathcal{P}(x)\right]^{-1}, \tag{6}$$

i.e., the replication speed is inversely proportional to the logarithmic slope of $\mathcal{P}(x)$ [22, 23]. An advantage of the deterministic assumption is that it leads to a one-to-one correspondence between DNA abundance and local replisome speed thanks to Eq. (6). However, in many realistic cases, neglecting stochasticity might lead to inaccurate predictions.

Unfortunately, in the general stochastic case, different replication programs might give rise to the same DNA abundance distribution. This implies that one can not directly invert Eq. (3) to express $f(x, \tau)$ in terms of $\mathcal{P}(x)$. In these situations, one has to complement the information contained in $\mathcal{P}(x)$ with modeling assumptions, as exemplified in the following.

III. BACTERIAL DNA REPLICATION

A. Model

In this Section, we introduce a stochastic model for the DNA replication of the bacterium *E. coli*. The model accounts for variations in the speed of replication, as recently observed [21]. In addition, replisomes in the model can stochastically stall, as observed in single molecule experiments [28, 29].

Most bacteria, including *E. coli*, have a single circular chromosome that is replicated by two replisomes. The two replisomes start from the same origin site, one proceeding clockwise and the other counterclockwise. Replication is completed when the two replisomes meet. We assume that the two replisomes do not backtrack and that they act independently from one another until they meet. A base is therefore replicated by whichever replisome reaches it first. The joint replication program of two replisomes is therefore $f(x, \tau) = 1 - [1 - f_1(x, \tau)][1 - f_2(x, \tau)]$, where $f_1(x, \tau)$ and $f_2(x, \tau)$ are their individual replication programs, i.e. the probabilities that position x has been replicated at age τ by the respective replisome. Substituting this expression in Eq. (2) we obtain

$$\mathcal{P}(x) = 1 - \int_0^\infty d\tau \Lambda e^{-\Lambda\tau}\left(1 - f_1(x, \tau)\right)\left(1 - f_2(x, \tau)\right). \tag{7}$$

We define the genome coordinate $x \in [0, L]$ where L is the genome length. We set the coordinate of the replication origin at $x = 0$ (and equivalently $x = L$, since the genome is circular). We call $x_1(\tau)$ and $x_2(\tau)$ the positions of the two replisomes along this coordinate as a function of age. The first replisome starts at $x(0) = 0$ and moves in the direction of increasing x, whereas the second starts at $x(0) = L$ and moves in the direction of decreasing x.

We express the replisome dynamics in terms of two

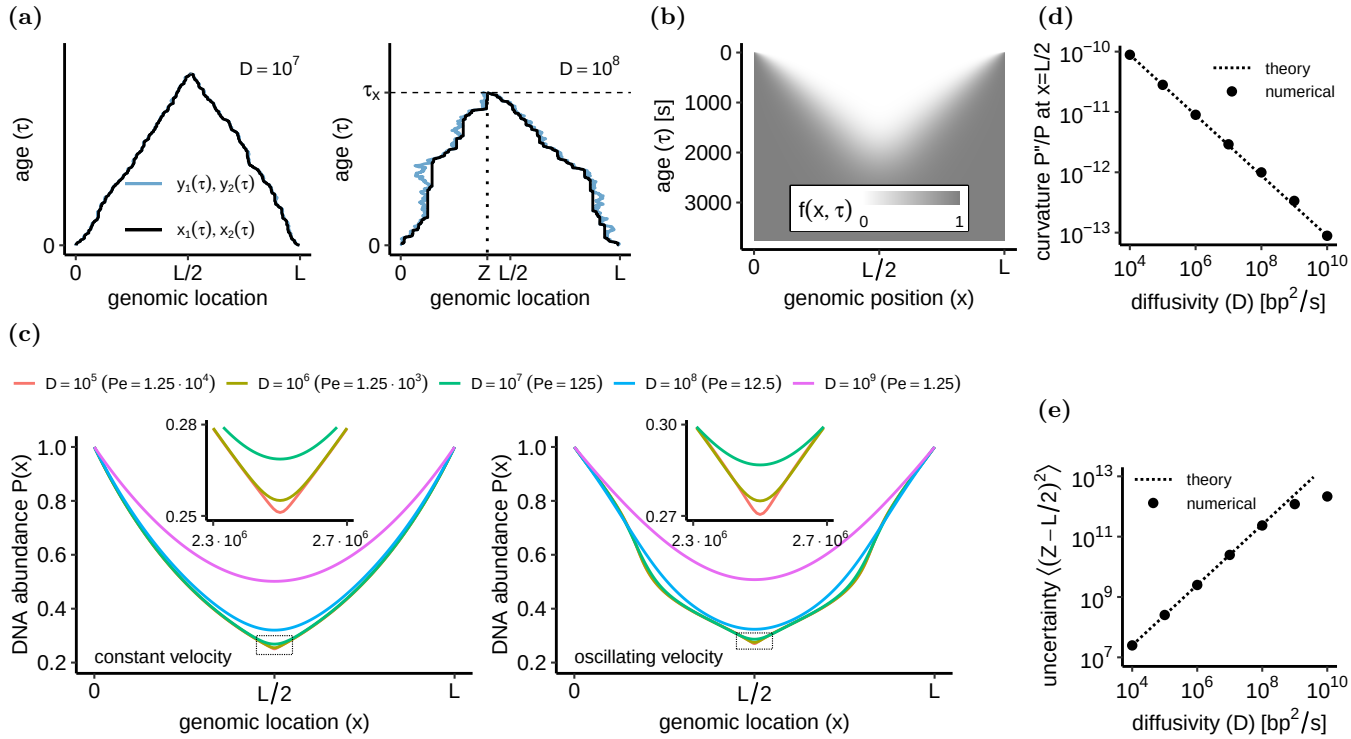

FIG. 2. Bacterial DNA replication model. Parameters are: genome length $L = 5 \cdot 10^6$, growth rate $\Lambda = 2\text{h}^{-1}$, baseline speed $v_0 = 10^3$ bp/s. **(a)**. Trajectories x_1, x_2 of the model (black lines) and auxiliary processes y_1, y_2 (blue lines). Backwards movements of y_1, y_2 correspond to stochastic stalling of x_1, x_2. Replication concludes at an age τ_x when x_1 and x_2 first meet at a random meeting point $Z = x_1(\tau_x) = x_2(\tau_x)$. **(b)**. Bacterial replication program $f(x,\tau) = 1 - (1 - f_1)(1 - f_2)$ with f_1, f_2 from Eq. (9) for $D = 10^8$ bp^2/s. **(c)**. DNA abundance distribution for constant and oscillating replisome speed and different values of D. In the constant speed case, $h(\tau) = 1$, whereas in the oscillating speed case $h(\tau) = 1 + \delta\cos(\omega\tau + \phi)$ with $\delta = 0.5$, $\omega = 2\pi/1800$, and $\phi = 0$. **(d)**. Curvature $\mathcal{P}''(x)/\mathcal{P}(x)$ at the terminus $x = L/2$ predicted by Eq. (11) vs. numerical simulations. **(e)**. Uncertainty of the meeting point Z predicted by Eq. (12) vs. numerical simulations.

Langevin equations

$$\begin{aligned}
\frac{d}{d\tau}y_1 &= v_0 h(\tau) + \sqrt{2Dh(\tau)}\,\xi_1(\tau) \\
\frac{d}{d\tau}y_2 &= -v_0 h(\tau) + \sqrt{2Dh(\tau)}\,\xi_2(\tau) \\
x_1(\tau) &= \max_{\tau'\le\tau} y_1(\tau') \\
x_2(\tau) &= \min_{\tau'\le\tau} y_2(\tau')\ ,
\end{aligned} \tag{8}$$

where $\xi_1(\tau)$ and $\xi_2(\tau)$ are Gaussian white noise sources satisfying $\langle \xi_i(\tau)\xi_j(\tau')\rangle = \delta_{ij}\delta(\tau-\tau')$. In Eq. (8), we have assumed that the instantaneous average speeds of the two replisomes are opposite in sign. These instantaneous speeds are modulated in time by a function $h(\tau)$, as result of varying conditions during the cell cycle (see [21]). We have assumed that the noise amplitude $\sqrt{2Dh(\tau)}$ is modulated in time by same the function $h(\tau)$ as the instantaneous average speed. This assumption is also convenient to mathematically solve the model.

The physics of Eq. (8) can be interpreted as follows. Whenever $y_i(\tau)$ attains a new maximal distance from its origin, then $x_i(\tau)$ is moving forward and it coincides with $y_i(\tau)$. If, instead, $y_i(\tau)$ is making a negative excursion from its past maximal distance, then $x_i(\tau)$ remains frozen at the value of the last maximal distance of $y_i(\tau)$. We interpret these stochastic pausing as replisome stalling events. The amplitude of the noise terms in Eqs. (8) controls the commonness of long stalling events. Stochastic trajectories generated by the model are qualitatively similar of those observed in single-molecule experiments with DNA polymerases [28, 29], see Fig. 2a.

The balance between replisome progress and stalling is governed by the dimensionless Peclet number $\text{Pe} = Lv_0/4D$. For large Pe, the dynamics are nearly deterministic and long stalling events are infrequent.

A consequence of Eq. (8) is that the individual replication program $f_i(x,\tau)$ is equal to the first-passage probability of the associated process y_i through x. This first-passage probability is expressed by

$$f_i(x,\tau) = \int_0^{H(\tau)} \sqrt{\frac{\mu_i^3}{2\pi\sigma_i^2 u^3}}\,\exp\left(-\frac{1}{2}\frac{(u-\mu_i)^2}{\sigma_i^2}\frac{\mu_i}{u}\right)du\,, \tag{9}$$

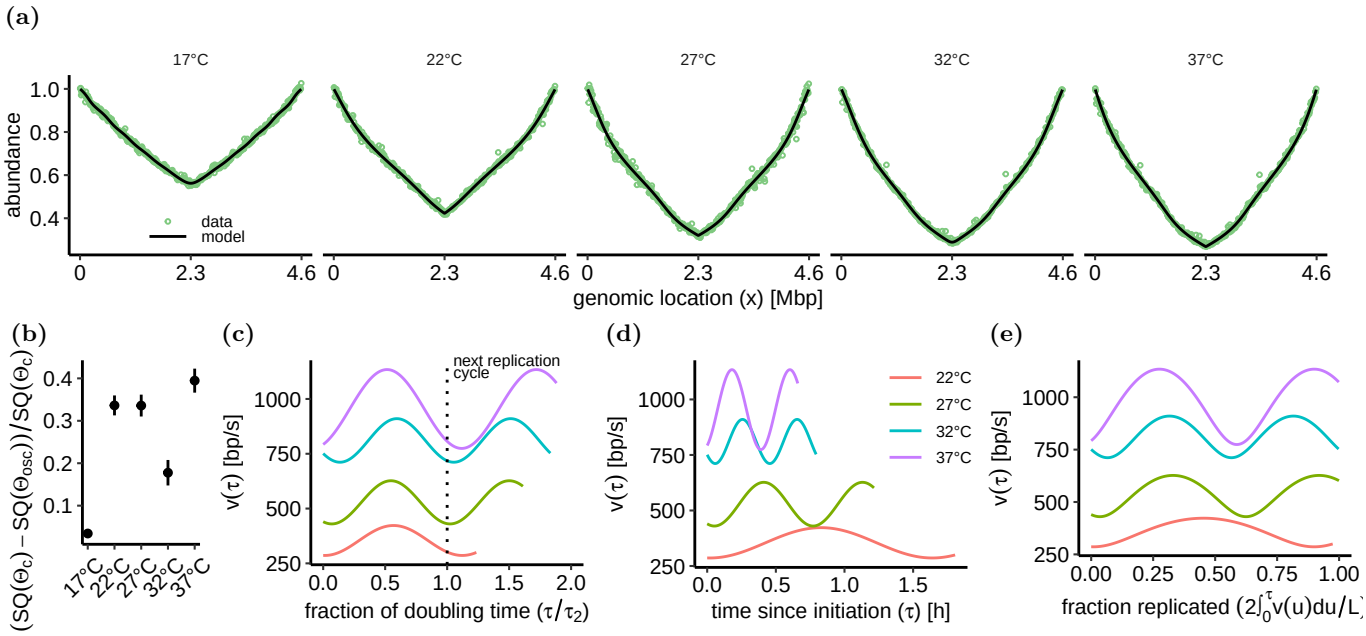

FIG. 3. Bacterial model fitted to the data of Ref. [21]. The replisome speed is modulated by an oscillatory function, see Eq. (13). We fitted the parameters v_0, D, δ, ω, and ϕ from the measured DNA abundances. The growth rate Λ was independently measured in the experiments (see Appendix E). **(a)**. Observed DNA abundance and model predictions for E. coli cultures growing at temperatures $T = 17°C, 22°C, 27°C, 32°C, 37°C$. **(b)**. Relative decrease in residuals ($\mathrm{SQ}(\Theta) = \sum_{i=1}^{N}(a_i - \lambda\mathcal{P}(x_i|\Theta))^2/\sigma_i^2$, see Eq. (E2)) of the model (Θ_{osc}) vs. the constant speed case (Θ_{c}). **(c)** Average instantaneous speed $v(\tau) = v_0 h(\tau)$ as a function of the fraction τ/τ_2 of the doubling time $\tau_2 = \log(2)/\Lambda$. **(d)**. Average instantaneous speed $v(\tau) = v_0 h(\tau)$ as the function of the time τ since replication initiation.**(e)** Average instantaneous speed $v(\tau) = v_0 h(\tau)$ as a function of the replication progress, i.e. of the fraction $2\int_0^\tau v(u)du/L$ of replicated genome. In (c-e) we omitted 17°C since the effect of speed fluctuations on $\mathcal{P}(x)$ is negligible at that temperature (see panel b).

where $H(\tau) = \int_0^\tau h(u)du$ and

$$\mu_1 = x/v_0, \qquad \mu_2 = (L-x)/v_0,$$
$$\sigma_1^2 = 2Dx/v_0^3, \qquad \sigma_2^2 = 2D(L-x)/v_0^3 \ . \tag{10}$$

Eqs. (9) and (10) are derived in Appendix A. We compute the DNA abundance $\mathcal{P}(x)$ by substituting Eq. (9) into Eq. (7), see Fig. 2b. We numerically evaluate the final integral over τ appearing in Eq. (7).

The main effect of diffusivity is to smoothen the DNA abundance distribution predicted by the model, see Fig. 2c. This effect is particularly pronounced at the expected meeting point $x = L/2$ of the two replisomes, where the DNA abundance exhibits a cusp for $D = 0$ but is smooth for positive D, see Figure 2c. In contrast, far from the meeting point diffusivity does not affect the DNA abundance much, see Appendix B. We now quantify the effect of diffusivity by means of Eq. (9). In the case of constant speed $v(\tau) = v_0$, the relative curvature of $\mathcal{P}(x)$ at the expected meeting point is approximated by

$$\frac{\mathcal{P}''(L/2)}{\mathcal{P}(L/2)} \approx \frac{2\Lambda}{\sqrt{\pi D L v_0}} \ , \tag{11}$$

see Fig. 2d and Appendix B. In principle, Eq. (11) can be used to estimate the diffusion coefficient from the curvature in the terminus region, without having to fit the entire distribution.

From the point of view of trajectories, the smoothing of $\mathcal{P}(x)$ for $D > 0$ occurs because the two replisomes no longer necessarily meet exactly at $x = L/2$. In particular, we find that the uncertainty on the location Z of the meeting point is approximately equal to

$$\left\langle (Z - L/2)^2 \right\rangle \approx \frac{DL}{2v_0} = \frac{L^2}{8\,\mathrm{Pe}} \ , \tag{12}$$

see Fig. 2e. Equation (12) is derived in Appendix C.

B. Comparison with E.coli sequencing data

We now fit our model of replisome dynamics to experimentally measured DNA abundance from wild type *E. coli* grown at different temperatures (from Ref. [21]). We assume that the speed and diffusion coefficient are modulated in time by a function

$$h(\tau) = 1 + \delta\cos(\omega\tau + \phi), \tag{13}$$

see Fig. 2. In the fit, we treat v_0, D, δ, ω, and ϕ as free parameters (see Appendix E).

The model fits the experimental DNA abundances very well, see Fig. 3a. The estimates of the mean speed v_0 are

T [°C]	v_0 [bp/s]	D [kbp^2/s]	δ	ω [rad/h]	ϕ [rad]	Pe	ΔZ [kbp]
17	230 ± 30	1.6 ± 1.4	0.29 ± 0.1	5.8 ± 4.2	1.3 ± 1.5	100 ± 100	130 ± 50
22	350 ± 21	0.5 ± 0.8	0.19 ± 0.07	3.9 ± 0.6	3.1 ± 0.5	200 ± 30	110 ± 7
27	530 ± 20	1.0 ± 1.2	0.19 ± 0.03	8.7 ± 0.4	2.0 ± 0.2	400 ± 400	80 ± 40
32	810 ± 57	1.7 ± 1.8	0.12 ± 0.04	16 ± 1.3	1.5 ± 0.2	500 ± 400	80 ± 30
37	950 ± 24	1.2 ± 2.3	0.19 ± 0.03	15 ± 0.4	2.9 ± 0.2	500 ± 600	70 ± 50

TABLE I. Parameter estimates for time-dependent speed $v(t) = v_0\big(1 + \delta\cos(\omega t + \phi)\big)$. For v_0, D, δ, ω, ϕ the reported standard errors represent the variability over replicates. The Peclet number $Pe = Lv_0/4D$ and meeting point uncertainty $\Delta Z = \sqrt{DL/2v_0}$, see Eq. (12), are computed from the average estimates of v_0 and D over replicates where $D > 0$. Their standard error are estimated using error propagation.

highly consistent among replicates across all temperatures, see Table I. At temperatures above 17°C we find robust evidence of speed fluctuations. The model provides consistent estimates of δ, ω and ϕ in these cases, with an improvement of the quality of fit ranging between 20% to 40% percent compared to the constant-speed case (Fig. 3b). At 17°C, the effect of the speed fluctuations on $\mathcal{P}(x)$ is small and, consequently, the uncertainty in the associated parameters ω and ϕ is high. Regardless of temperature, model selection appears to prefer a vanishing value of D for some replicates. The reason is likely that the estimates of D correspond to Peclet numbers in the range from Pe ≈ 200 to Pe ≈ 1000, which lies close to the detection threshold; see Appendix E.

The frequency of speed oscillations appears linked with the population doubling time. In fact, the oscillations for 22°C to 37°C align well when time is rescaled according to the duration of a cell cycle, see Fig. 3c. The speed consistently attains a minimum after one doubling time $\tau_2 = \log(2)/\Lambda$ when the next replication cycle starts and the number of active forks is thus increased. As comparisons, when plotting the oscillations against absolute time since replication initiation (Fig. 3d) or against replication progress (Fig. 3e) the alignment is substantially worse.

The speed oscillations were first observed and quantified using a model in which speed is modulated in space, rather than in time [21]. Our approach permits to analytically solve also a spatially modulated speed model in the small-noise limit. The resulting parameter values well match those of Ref. [21], and are consistent with our temporally modulated model, see Appendix D for details. The model with temporal speed oscillations yields a better fit for the majority of samples, consistently with the idea that the speed variations are linked with the doubling time. The improvement in likelihood is, however, small (Fig. 7b in Appendix D).

IV. EUKARYOTIC DNA REPLICATION

A. Model

In this Section, we apply our approach to eukaryotic DNA replication. The eukaryotic case is substantially more complex than the bacterial one, because of the presence of multiple, randomly activated replication origins. Also in this case, when an origin fires, two replisomes start moving from that origin in opposite directions. In this case, theoretical progress [30, 31] has made use of an analogy with the physics of freezing/crystallization kinetics, as described by the so-called Kolmogorov-Johnson-Mehl-Avrami model [32–34], see also [35, 36]. We briefly summarize this idea and then extend it to asynchronously growing populations.

To fix the ideas, we consider a given location x at an age τ in a eukaryotic genome. A main problem of the eukaryotic case is that, in principle, a location x could have been replicated by different replisome starting from different origins. The replication program $f(x,\tau)$ can be seen as the probability that the past "light-cone" $V_{x,\tau}$ of the space-time point x,τ contains at least one origin firing event, see Fig. 4a. This elegant argument circumvents the problem of determining from which origin did the replisome that replicated the genome location x start.

Assuming that replisomes progress deterministically with constant speed v_0, the past light-cone is the set of space-time points x',τ' from which x is reachable by a replisome within a time τ, i.e.,

$$V_{x,\tau} = \big\{(x',\tau')\,\big|\,|x-x'| \le v_0\,(\tau-\tau')\,\big\}, \qquad (14)$$

see Fig. 4a. Following [1, 37], we now assume that origins fire independently from one another in space and time with position- and age-dependent rates $I(x,\tau)$. This assumption implies that the space-time points at which origins fire form a two-dimensional Poisson point process.

The probability that location x has been replicated by age τ is then expressed by

$$f(x,\tau) = 1 - \exp\left(-\iint_{V_{x,\tau}} dx'd\tau'\, I(x',\tau')\right). \qquad (15)$$

We note that Eq. (15) would remain valid if we chose a different replisome dynamics, corresponding to a form of the light-cone $V_{x,\tau}$ other than that expressed by Eq. (14). Using Eq. (2), we formally express the DNA abundance distribution as

$$\mathcal{P}(x) = 1 - \Lambda\int_0^\infty d\tau \exp\left(-\Lambda\tau - \iint_{V_{x,\tau}} dx'd\tau'\, I(x',\tau')\right). \qquad (16)$$

We now focus on budding yeast, where origins correlate with specific DNA motifs and are therefore thought to be

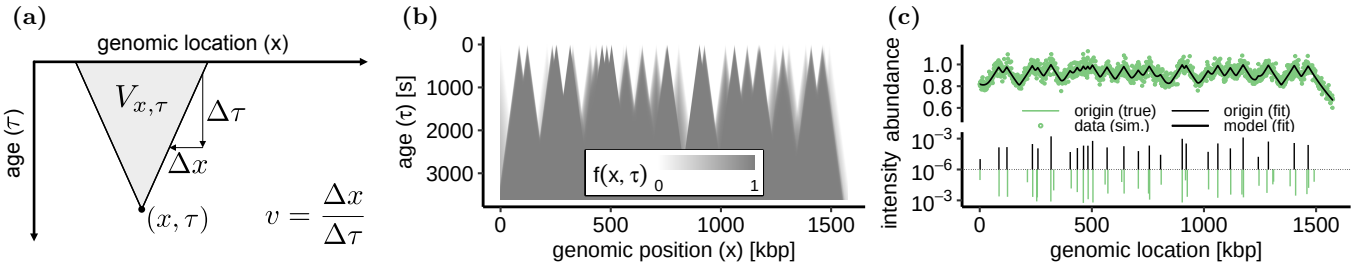

FIG. 4. Eukaryotic model. **(a)**. Past light-cone $V_{x,\tau}$ of a space-time point (x,τ). At least one origin must have fired within the light cone for location x to be replicated by time τ. **(b)**. Eukaryotic replication program from Eq. (21) for the annotated origins on *S. cerevisiae* W303 chromosome IV with intensities $I_i^\star/v$ randomly log-uniformly distributed on $[10^{-5}, 2\cdot 10^{-}3]$. **(c)**. Inference of origin locations and intensities from a simulated DNA abundance. The top plot shows the model (black line) fitted to simulated abundances (green line). The bottom plot shows the inferred 26 origins (black bars) and 39 true origins (green bars). Rescaled ntensities $I_i^\star/v$ [1/bp] are plotted in log-scale. Parameters are: $v = 27$ bp/s, $\Lambda = 9.6\cdot 10^{-5}$ 1/s (doubling time 120 minutes). The true origin locations and intensities are those from panel (b). The resulting true abundances are multiplied by a Gamma-distributed random variable with mean 1.0 and coefficient of variation 0.04 to mimic measurement errors.

at well-defined locations $x_1,\ldots,x_K$ on the chromosomes [6]. We therefore assume $I(x,\tau) = \sum_{j=1}^K I_j(\tau)\delta(x-x_j)$. We further assume that firing rates are constant in time, so that

$$I(x,\tau) = \sum_{j=1}^K I_j^\star \delta(x-x_j). \tag{17}$$

In budding yeast, the origin firing rate was observed to be time-dependent [8]. This means that this second assumption is not fully accurate and should be considered as a simplifying approximation.

To express the DNA abundance $\mathcal{P}(x)$ for firing rates as in Eq. 17, we consider the travel time τ_k from origin x_k to a given position x,

$$\tau_k = \frac{|x-x_k|}{v_0}. \tag{18}$$

At fixed x, we reorder the origins so that $0<\tau_1<\cdots<\tau_K$. We then define the effective replication times

$$\mathcal{T}_k = \tau_k + \sum_{j=1}^k (\tau_k-\tau_j)\, I_j^\star/\Lambda \tag{19}$$

which collectively take into account the effect of all origins. We also introduce the effective cumulative weights

$$W_k = 1 + \sum_{j=1}^k I_j^\star/\Lambda, \tag{20}$$

with the convention that $W_0 = 1$. Using these definitions, we express the DNA abundance by

$$\mathcal{P}(x) = \sum_{k=1}^K \left(\frac{1}{W_{k-1}} - \frac{1}{W_k}\right) e^{-\Lambda\mathcal{T}_k}. \tag{21}$$

Equation (21) is derived in Appendix F.

We implemented a simulated annealing algorithm to infer the origin number, location, and intensities based on DNA abundance data via Eq. (21). Our inference procedure treats Λ/v_0, the number of origins K, the origin positions $x_1,\ldots,x_K$, and the re-scaled firing rates $I_1^\star/v_0,\ldots,I_K^\star/v_0$ as free parameters. We call the compound parameter $I_j^\star/v_0$ the intensity of origin j. Our algorithm uses the Akaike information criterion (AIC) as the cost function to avoid over-fitting (details in Appendix G).

We first tested this inference procedure on artificial data, in which the genome length (1.6Mbp) and origins distribution are comparable to those of the longest chromosome of budding yeast (chromosome IV). Our inference algorithm detected the location and intensity of 26 out of 39 true origins with high accuracy (Fig. 4c; median distance to true origin 1.5kb, median relative error of intensity 40% if non-resolved clusters of true origins are merged; true intensities range over two orders of magnitude). The 13 non-recovered origins either have low intensity or were merged with another origin in close proximity.

B. Application to experimental data

We now apply our model and inference procedure to experimentally measured DNA abundance in an asynchronous populations of budding yeast (*S. cerevisiae* W303) [38]. Our algorithm infers $K = 234$ origins across the 16 yeast chromosomes. The predicted DNA abundance well matches the experimental one, see Fig. 5a.

Our method well predicts origins of replication with a resolution on the order of a few kilobases, and without using additional experimental information other than the DNA abundance distribution. The inferred origins locations correlate with known origins of *S. cerevisiae* when binned on a length scale of about 7kb, see Fig. 5b. A second correlation peak at 100kb suggests a large-scale pattern in the distribution of origins. As expected, the peak at 7kb vanishes when known origin locations are randomly shifted by ±35kbp, or shuffled by randomly

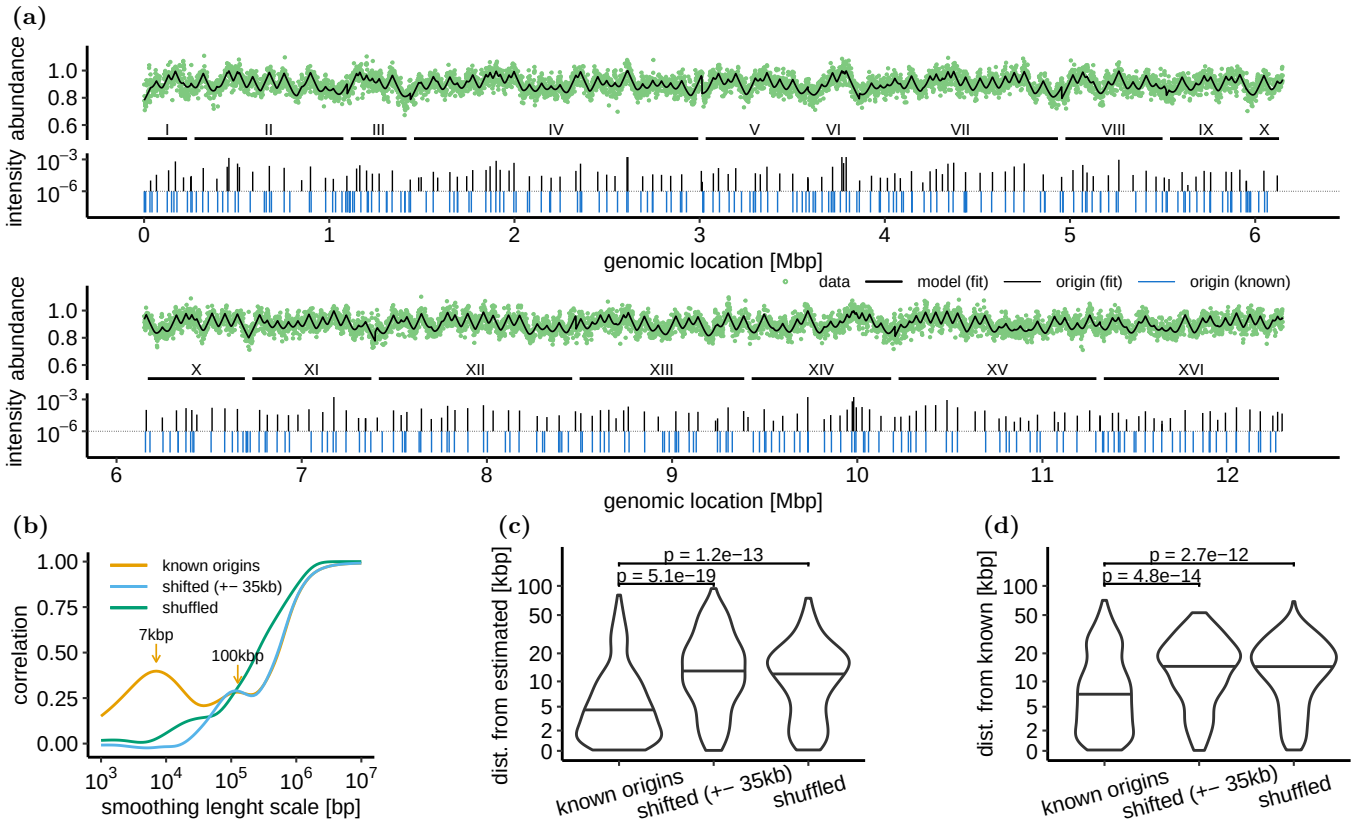

FIG. 5. Eukaryotic model fitted to *S. cerevisiae* data from [38]. Fitted parameters are: Λ/v, origin positions $x_1, \ldots, x_n$ and intensities $I_1^\star/v, \ldots, I_n^\star/v$. The fitted number of origins is $K = 234$. Known origins used for validation were lifted from the annotated genome for strain 288C (RefSeq assembly *R64-3-1*, accession *GCF_000146045.2*) using *liftoff* [39]. We excluded the mitochondrial genome. **(a)**. Observed (green circles) and predicted (black line) DNA abundances (top) and inferred origin positions and intensities (bottom). Known origin positions (blue) shown for comparison. **(b)**. Correlation of estimated ($K = 234$) and known ($K = 354$) origin densities at different smoothing length scales (yellow). For comparison, we also show the correlation to the density of shuffled (green) and randomly shifted (blue) known origins. **(c)**. Distribution of distances between estimated and closest known origins. **(d)**. Distribution of distances between known and closest estimated origins. In (c) and (d), the y-axis is non-linearly scaled to enlarge the region of interest. Horizontal lines indicate the medians. Reported p-values are computed using Mann–Whitney U tests.

reordering chromosomes. The results of the correlation analysis are corroborated by matching each inferred origin to the closed known origin (figure 5c; median distance 3.9kb) respectively each known origin to the closest inferred origin (figure 5d; median distance 6.6kb). In both cases, the average distances are significantly increased if the known origins are shifted or shuffled.

V. EXTENSIONS OF THE THEORY

A. Non-exponential growth

Our approach can be generalized to populations that do not grow exponentially. We here assume that the number of genomes grows with time according to an arbitrary function $N_g(t)$. In this case, following the same logic of Section II, we find that Eq. (2) becomes

$$\mathcal{P}(x,t) = \int_0^\infty d\tau P(\tau;t) f(x,\tau) \tag{22}$$

where $P(\tau;t) \propto N_g'(t-\tau)$. The proportionality constant can be determined by imposing normalization of $P(\tau)$. After integrating Eq. 22 by parts, we obtain

$$\mathcal{P}(x,t) = \langle N_g(t-\tau) \rangle \,. \tag{23}$$

This generalization can be applied, for example, to bacterial or eukaryotic populations that grow in a partially synchronous manner. In such cases, $N_g(t)$ might exhibit temporal peaks.

We note that, in Eqs. (22) and (23) we implicitly made the assumption that the replication program $f(x,\tau)$ does not depend on t. The validity of this assumption should be verified for populations that are out of steady growth. Moreover, a practical limiting factor can be the accuracy at which $N_g(t)$ can be estimated.

B. Non-immortal genomes

For the purpose of computing the DNA abundance distribution, our assumption of immortal genomes is justified as long as typical genome lifetimes $\tau_{\max}$ are much longer than the doubling time $\tau_2 = \log(2)/\Lambda$. This is usually a realistic assumption, but in some cases one may want to generalize our approach to include genome degradation. In such case, the exponential distribution in Eq. (2) should be replaced with the age distribution $P(\tau)$ of surviving genomes. In particular, if genomes survive until age τ with probability $s(\tau)$, the age distribution is expressed by

$$P(\tau) \propto \Lambda e^{-\Lambda\tau} s(\tau) \tag{24}$$

up to a normalization constant.

C. Multiple genome types

The theory we developed so far applies to statistically identical genomes. This assumption is sometimes referred to as "identity at birth" in the literature on renewal processes [40]: all genomes are born identical, and the differences that they might develop as they synthesize can be seen as independent outcomes of the same stochastic process.

There might be cases in which this assumption does not hold. For example, certain knockout mutations of *E. coli* lead to defective genome types that are inheritable [41]. To address these scenarios, we extend our theory to cases in which different genome types coexist in the population, each characterized by a different replication program.

We call $N_g^{(i)}(t)$ the number of genomes of type i at time t in the population, with $i = 1 \dots k$. The total number of genomes is $N_g(t) = \sum_i N_g^{(i)}(t)$. Each genome of type i can be replicated into a genome of type i' according a certain stochastic process. We assume that this stochastic process satisfies the Perron-Frobenius theorem, so that at long time the number of genomes are given by

$$N_g^{(i)}(t) \propto r_i\, e^{\Lambda t}. \tag{25}$$

Here Λ is the leading eigenvalue of the dynamics, corresponding to the exponential growth rate of the population. The vector $(r_1 \dots r_k)$ is the leading eigenvector, having non-negative real entries according to the Perron-Frobenius theorem, and normalized such that $\sum_i r_i = 1$. The resulting replication program can then be obtained as an average over the genome types:

$$f(x,\tau) = \sum_{i=1}^{k} r_i f^{(i)}(x,\tau). \tag{26}$$

This average replication program can be substituted into Eq. (2) to determine the DNA abundance distribution.

D. Numerical simulations and analogy with stochastic resetting

The models we considered in Sections III and IV are analytically solvable. However, this might not be the case for more complex models. In this Section, we briefly present an interpretation of the DNA abundance distribution $\mathcal{P}(x)$ as the steady-state of a dynamical process implementing stochastic resetting [42, 43]. This analogy can be used for efficient numerical simulations.

The models introduced in Sections III and IV are described by linear equations. Formally, we assume that we can always write the stochastic process associated with a certain DNA replication program in terms of linear evolution operator $\hat{\mathcal{L}}$, so that

$$\partial_\tau f(x,\tau) = \hat{\mathcal{L}} f(x,\tau). \tag{27}$$

We now introduce a modified version of this equation:

$$\partial_\tau f(x,\tau) = \hat{\mathcal{L}} f(x,\tau) - \Lambda f(x,\tau). \tag{28}$$

Equation 28 can be interpreted as describing a replication program equivalent to that of Eq. (27), but that is stochastically reset to its beginning at rate Λ. Stochastic simulations of this process can be easily implemented numerically.

We note that $\mathcal{P}(x)$ is a steady solution of Eq. (28), i.e., it satisfies

$$\left(\hat{\mathcal{L}} - \Lambda\right)\mathcal{P} = 0. \tag{29}$$

This means that we can determine $\mathcal{P}(x)$, up to an appropriate normalization constant, by sampling the stochastic dynamics described Eq. (28) at steady state. Such stochastic simulations can be efficiently implemented numerically [21].

VI. CONCLUSIONS

In this paper, we introduced a general theory that connects the DNA replication program with the abundance of DNA fragments that one should expect in an asynchronously growing population of cells. Our theory builds on previous approaches [3, 15, 21, 22, 25] and has the advantage of being based on a minimal set of realistic assumptions and allowing for stochastic replication programs. As we have demonstrated, these key properties make our theory applicable to a broad range of organisms, from bacteria to eukaryotes.

In the case of bacterial DNA replication, we proposed a model in which the replisome speed is modulated in time and replisomes can stochastically stall. We solved the model exactly and fitted its prediction against sequencing data of *E.coli* growing at different temperatures [21]. The fits show that the period of speed oscillations matches the population doubling time, or equivalently the time interval between consecutive fork firing. Our model with

time-periodic speed variations fits the data slightly better than the one with space-periodic variations as postulated in [21]. Taken together, these observations support that the causes of oscillations are linked with the cell cycle, or alternatively with the fork firing rate. A possible candidate would be competition among multiple forks on the same genome [21]. At variance with Ref. [21], the approach introduced in this paper leads to an analytical expression for the DNA abundance distribution, which considerably simplifies the inference procedure and provides additional physical insight.

The fits of the *E.coli* data reveal that the Peclet number characterizing the replisome dynamics is rather large. On the one hand, this observation confirms that simpler approaches that neglect stochasticity [22, 23] provide reliable results, at least in the case of wild type *E.coli*. On the other hand, the diffusion constant, albeit small, provides important information about the uncertainty of the replisome meeting point. In *E.coli*, the Tus-Ter system is know to set bounds on the region in which replisomes can meet [44], thereby likely affecting this accuracy. It will be interesting for future studies to apply our approach to mutant strains, to see whether they are characterized by a different degree of uncertainty.

Stochasticity is definitely crucial in the eukaryotic case. We have used our approach to estimate the origins location and intensities for budding yeast from the DNA abundance distribution measured in [45]. Our approach is based on seminal work by Bechhoefer and coworkers [1, 37], that we extended to asynchronously growing populations. A previous study [3] also attempted at extending the approach from [1, 37] to asynchronously growing budding yeast. Our results differ from those of Ref. [3] in two different aspects. First, Ref. [3] assumed as a working hypothesis a uniform distribution for the distribution $P(\tau)$. In contrast, our central result in Section II shows that the distribution $P(\tau)$ should be exponential under very general conditions. Second, in fitting the model to the data, Ref. [3] used experimental knowledge of the origin coordinates. Instead, our method was able to directly infer these coordinates, without requiring any additional experimental information other than the DNA abundance distribution.

In our eukaryotic model, we assumed for simplicity that replisome speed is constant; that origins are placed at well defined sites; and that they fire at an origin-dependent rate that is constant in time. The last assumption, in particular, is a drastic approximation, since origin firing rates in yeast are known to be markedly time-dependent [8]. Relaxing these assumptions constitutes an important challenge for future research and will permit to recover origin timing behaviour, beside locations, and thus provide a more complete picture of the replication program.

In any case, despite these simplifying assumptions, our algorithm successfully recovers the locations of the majority of known origins in budding yeast, with an accuracy on the order of kilobases. The accuracy can likely be further increased by exploiting advances in sequencing technology, in particular increased sequencing depth and read lengths, and by further improving the optimization algorithm. Our results demonstrate that the combination of deep sequencing of asynchronous populations and our inference approach provides a cost-effective way of discovering the replication origins of any single-cell eukaryotic species which can be cultured and sequenced.

ACKNOWLEDGMENTS

We thank S. Hauf, C. Plessy, and Y. Yokobayashi for fruitful discussions. We thank A. Alsina, J. Bechoefer, N. Rhind, P. Sartori, and A. Sassi for feedback on a preliminary manuscript. This work was supported by JSPS KAKENHI Grant No. 23H01146 (to SP). DB thanks Vellore Institute of Technology, Vellore for providing VIT SEED Grant-RGEMS Fund (SG20220060) for carrying out this research work.

Appendix A: Bacterial DNA replication with time-dependent speed

We consider the Langevin Eqs. (8). We write the associated Fokker-Planck equation for the first replisome:

$$\frac{\partial}{\partial\tau}p = h(\tau)\left(-v_0\frac{\partial}{\partial y}p + D\frac{\partial^2}{\partial y^2}p\right). \tag{A1}$$

Equation (A1) shows that the function $h(\tau)$ acts as global time rescaling. Therefore, if $\hat{p}$ is a solution of the Fokker-Planck equation with constant drift and diffusion terms

$$\frac{\partial}{\partial\tau}\hat{p} = -v_0\frac{\partial}{\partial y}\hat{p} + D\frac{\partial^2}{\partial y^2}\hat{p} \tag{A2}$$

under certain boundary and initial conditions, then

$$p(y,\tau) = \hat{p}\big(y, H(\tau)\big) \tag{A3}$$

solves the original Fokker-Planck Eq. (A1) with the same boundary and initial conditions, where $H(\tau) = \int_0^\tau du\, h(u)$. We impose, in particular, an absorbing boundary condition at x, $\hat{p}(x,\tau) = 0\ \forall\tau$ and an initial condition $p(y,\tau=0) = \delta(y)$. The flux $\hat{J} = v_0 p - D\partial_y p$ through the absorbing boundary is given by an inverse Gaussian distribution:

$$\hat{J}(x,\tau) = \frac{x}{\sqrt{4\pi D\tau^3}}e^{-\frac{(x-v_0\tau)^2}{4D\tau}}. \tag{A4}$$

Since the first-passage density of Eq. (A1) equals the flux $J(x,\tau) = h(\tau)\hat{J}\big(x, H(\tau)\big)$, we have

$$\frac{\partial}{\partial\tau}f_1(x,\tau) = h(\tau)\hat{J}\big(x,H(\tau)\big) = \frac{h(\tau)x}{\sqrt{4\pi DH^3(\tau)}}e^{-\frac{(x-v_0H(\tau))^2}{4DH(\tau)}}. \tag{A5}$$

The same approach can be used for the second replisome after a suitable change of coordinates and initial conditions. Integrating Eq. (A5) over τ, and similarly for the second replisome, directly leads to Eqs. (9) and (10).

Appendix B: Effect of diffusivity on the shape of the bacterial DNA abundance distribution

We here study the effect of increasing the diffusivity D on the shape of the DNA abundance distribution $\mathcal{P}(x)$. We first focus on the the meeting point region, and then consider the region far from the expected meeting point. We call the expected meeting point $x = L/2$ the "replication terminus". Since this is the last location on the genome to be replicated, $\mathcal{P}$ attains its global minimum there. For $D = 0$, $\mathcal{P}$ exhibits a cusp at the terminus, and thus infinite curvature. For $D > 0$, the cusp vanishes, the curvature is finite and decreases with D (Fig. 2c-d).

To quantitatively link D with this curvature, we first use that Eq. (8) is symmetric under the swap of replisomes and mirroring of the genomic coordinate: $f_1(x,\tau) = f_2(L-x,\tau)$. As a consequence, $f'(L/2)$ vanishes for all time τ and therefore $\mathcal{P}'(L/2) = 0$ according to Eq. 2. The curvature at $x = L/2$ is thus simply $\mathcal{P}''(L/2)$. We focus on the relative curvature expressed by

$$\frac{\mathcal{P}''(L/2)}{\mathcal{P}(L/2)} = \frac{\int_0^\infty d\tau \Lambda e^{-\Lambda\tau} f''(L/2,\tau)}{\int_0^\infty d\tau \Lambda e^{-\Lambda\tau} f(L/2,\tau)}. \tag{B1}$$

To make progress, we approximate the replication time distribution with a Gaussian:

$$\psi(x,\tau) = \frac{\partial}{\partial\tau} f(x,\tau) \approx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(t-\mu(x))^2}{2\sigma^2(x)}}, \tag{B2}$$

where the mean $\mu(x)$ and the variance $\sigma^2(x)$ have to be determined. It then follows from Eq. (3) that

$$\mathcal{P}(x) = \int_0^\infty d\tau e^{-\Lambda\tau}\psi(x,\tau) \approx e^{-\mu(x)\Lambda + \sigma^2(x)\Lambda^2/2}. \tag{B3}$$

Here, we have approximated the integral by extending the lower extreme to $-\infty$. We expect the error caused by this approximation to be negligible, since the Gaussian distribution that approximates $\psi(x,\tau)$ must be concentrated on the positive real numbers. By using that $f'(L/2,\tau) \equiv 0$ and thus $\mu'(x) = 0$, and assuming that the dependence of σ on x is small enough to be ignored, we obtain

$$\frac{\mathcal{P}''(L/2)}{\mathcal{P}(L/2)} \approx -\Lambda\mu''(L/2). \tag{B4}$$

We now have to evaluate $\mu''(L/2) = \partial_x^2 \int \tau\partial_\tau f(x,\tau)d\tau|_{x=L/2}$. To make this tractable, we now further approximate the individual inverse Gaussian laws of f_1, f_2 from Eq. (9) with Gaussian distributions. For these individual laws, the means and variances are given by Eq. (10) which yields

$$\hat{f}(x,\tau) = 1 - \left[1 - \Phi\left(\frac{\tau - x/v_0}{\sqrt{2Dx/v_0^3}}\right)\right] \times \left[1 - \Phi\left(\frac{\tau - (L-x)/v_0}{\sqrt{2D(L-x)/v_0^3}}\right)\right]. \tag{B5}$$

where we have for simplicity assumed that the replisome speed is constant, i.e. $h(\tau) = 1$. Therefore:

$$\begin{aligned}\mu''(L/2) &\approx \frac{\partial^2}{\partial x^2}\int_{-\infty}^{\infty} \tau\frac{\partial}{\partial\tau}\hat{f}(x,\tau)d\tau\Big|_{x=L/2} \\ &= \frac{\partial^2}{\partial x^2}\left(\int_0^\infty d\tau\big(1-\hat{f}(x,\tau)\big) - \int_{-\infty}^0 d\tau\hat{f}(x,\tau)\right)\Big|_{x=L/2} \\ &= -\int_{-\infty}^{\infty}\hat{f}''(L/2,\tau)d\tau \\ &= -\frac{2}{\sqrt{\pi D L v_0}}\end{aligned} \tag{B6}$$

and by inserting into Eq. (B4) we obtain Eq. (11) for the normalized curvature of the DNA abundance at the terminus.

Eq. B3 also permits to approximate the value of $\mathcal{P}(x)$ itself at the terminus. We use that

$$\begin{aligned}\mu(L/2) &\approx \int_{-\infty}^{\infty}\tau\frac{\partial}{\partial\tau}\hat{f}(x,\tau)d\tau \\ &= \int_0^\infty \big(1-\hat{f}(x,\tau)\big)d\tau - \int_{-\infty}^0 \hat{f}(x,\tau)d\tau \\ &= \frac{L}{2v_0} - \sqrt{\frac{DL}{\pi v_0^3}}\end{aligned} \tag{B7}$$

and

$$\begin{aligned}\sigma^2(L/2) &\approx \int_{-\infty}^{\infty}\tau^2\frac{\partial}{\partial\tau}\hat{f}(x,\tau)d\tau - \mu^2(L/2) \\ &= 2\int_0^\infty \tau\big(1-\hat{f}(x,\tau)\big)d\tau \\ &\quad - 2\int_{-\infty}^0 \tau\hat{f}(x,\tau)d\tau - \mu^2(L/2) \\ &= \frac{\pi-1}{\pi}\frac{DL}{v_0^3}.\end{aligned} \tag{B8}$$

Approximating $\mu(L/2+\epsilon) \approx \mu(L/2) + \epsilon^2\mu''(L/2)/2$ and $\sigma^2(L/2+\epsilon) \approx \sigma^2(L/2)$ in Eq. (B3) yields

$$\mathcal{P}(L/2+\epsilon) \approx \exp\left[-\frac{\Lambda L}{2v_0} + \Lambda\sqrt{\frac{DL}{\pi v_0^3}} + \frac{\pi-1}{\pi}\frac{\Lambda^2 DL}{2v_0^3} + \epsilon^2\frac{\Lambda}{\sqrt{\pi D L v_0}}\right], \tag{B9}$$

close to the replication terminus. The curvature at $x = L/2$ is therefore

$$\mathcal{P}''(L/2) \approx \frac{2\Lambda}{\sqrt{\pi D L v_0}}\exp\left[-\frac{\Lambda L}{2v_0} + \Lambda\sqrt{\frac{DL}{\pi v_0^3}} + \frac{\pi-1}{\pi}\frac{\Lambda^2 DL}{2v_0^3}\right]. \tag{B10}$$

We now study the shape of the DNA abundance far from the terminus region. Within these regions, only one of the two replisomes plays a relevant role. For simplicity, we focus on the region replicated by replisome 1. Also in this case, we assume constant replisome speed, i.e. $h(\tau) = 1$. Since replisome 2 has virtually no chance of reaching the area of interest before replisome 1, we assume $f(x,\tau) \approx f_1(x,\tau)$. The DNA abundance $\mathcal{P}(x)$ can then be explicitly found from the characteristic function of the inverse Gaussian distribution; the expansion $\sqrt{1+x} \approx 1 + x/2 - x^2/8$ yields for small D

$$\mathcal{P}(x) \approx \exp\left[\left(-\frac{\Lambda}{v_0} + D\frac{\Lambda^2}{v_0^3}\right)x\right]. \tag{B11}$$

The change in the exponential decay rate due to D is therefore of order $D\Lambda^2/v^3$.

Appendix C: Uncertainty on the meeting point

We now consider the random point Z at which the two replisomes meet and its uncertainty $\langle (Z - \langle Z\rangle)^2\rangle$. We call $\tau_1(x)$ and $\tau_2(x)$ the random times at which replisomes 1 and 2 reach location x, respectively. Then, $\tau_1(z) < \tau_2(z)$ for all points z to the left of the meeting point Z, and $\tau_1(z) > \tau_2(z)$ for all points to the right of Z. Therefore

$$\mathbb{P}(Z < z) = \mathbb{P}(\tau_1(z) > \tau_2(z)) \ . \tag{C1}$$

We note that the means and variances in Eq. (10) are the means and variances of $H\big(\tau_i(x)\big)$. Since $H(t)$ increases monotonically, we substitute it into Eq. (C1):

$$\mathbb{P}(Z < z) = \mathbb{P}\big(H(\tau_2(z)) - H(\tau_1(z)) < 0\big). \tag{C2}$$

By using that $\tau_1(z)$ and $\tau_2(z)$ are independent, and substituting the means and variances from Eq. (10), we find the mean and variance of $H(\tau_2(z)) - H(\tau_1(z))$ to be $(L-2z)/v_0$ and $2DL/v_0^3$, respectively. By approximating the distribution of the difference between the two replication times with a Gaussian, we obtain

$$\mathbb{P}(Z < z) \approx \Phi\left(\frac{-(L-2z)/v_0}{\sqrt{2DL/v_0^3}}\right) = \Phi\left(\frac{z - L/2}{\sqrt{DL/2v_0}}\right) \tag{C3}$$

where $\Phi(x) = 1/2 + \mathrm{erf}\left(x/\sqrt{2}\right)/2$ is the cumulative function of the Gaussian distribution. The uncertainly about the meeting point is therefore given by Eq. (12).

Appendix D: Bacterial DNA replication with position-dependent speed

In this Appendix, we solve a model with position-dependent speeds $v_1(x)$ and $v_2(x)$ in the limit of small diffusivity. This will allow us to directly compare time-dependent and position-dependent speed models. We allow the diffusion constants $D_1(x)$, $D_2(x)$ to vary in space as well. The Langevin equations are:

T	$\overline{v}$ [bp/s]	D [kbp^2/s]	δ	ω [rad/Mbp]	ϕ [rad]
17	230 ± 22	1.0 ± 1.2	0.18 ± 0.07	6.2 ± 6.7	2.3 ± 1.6
22	350 ± 37	1.0 ± 1.6	0.23 ± 0.15	2.8 ± 0.73	3.4 ± 0.72
27	540 ± 17	0.4 ± 0.5	0.18 ± 0.04	4.6 ± 0.11	2.1 ± 0.11
32	820 ± 57	1.4 ± 2.3	0.12 ± 0.04	5.5 ± 0.13	1.6 ± 0.16
37	970 ± 25	2.9 ± 2.3	0.18 ± 0.02	4.3 ± 0.15	3.0 ± 0.14

TABLE II. Parameter estimates for position-dependent speed $v(x) = \overline{v}\big(1 + \delta\cos(\omega x + \phi)\big)$ model of Bhat *et al.*[21].

$$\begin{aligned}\frac{d}{d\tau}x_1 &= v_1(x) + \sqrt{2D_1(x)}\,\xi_1(\tau)\\ \frac{d}{d\tau}x_2 &= v_2(x) + \sqrt{2D_2(x)}\,\xi_2(\tau).\end{aligned} \tag{D1}$$

where $v_1(x) > 0$ and $v_2(x) < 0$. As in the case of Eq. (8), the initial conditions are $x_1(0) = 0$, $x_2(0) = L$ and we define replication programs $f_1(x,\tau)$, $f_2(x,\tau)$ via the first-passage times of $x_1(\tau)$ and $x_2(\tau)$, respectively, through position x.

We now assume that (i) $v_1(x)$, $v_2(x)$ change only slowly, (ii) $D_1(x)$, $D_2(x)$ are small in comparison, and (iii) x is not too close to the replication origin. Under these assumptions, replication times are approximately additive, meaning that if $\tau_{x_0\to x_1}$ is the (random) time a replisome originating from x_0 takes to reach x_1, then $\tau_{x_0\to x_1}$ and $\tau_{x_1\to x_2}$ are independent, and $\tau_{x_0\to x_2} = \tau_{x_0\to x_1} + \tau_{x_1\to x_2}$. For large enough x, we invoke the central limit theorem for the sum $\tau_x = \tau_{x_0\to x_1} + \cdots + \tau_{x_k\to x}$ and conclude that ψ_1 and ψ_2 are approximately Gaussian, implying

$$f_i(x,\tau) \approx \frac{1}{2} + \frac{1}{2}\,\mathrm{erf}\left(\frac{\tau - \mu_i(x)}{\sqrt{2}\sigma_i(x)}\right) \tag{D2}$$

To find the position-dependent mean $\mu_i(x)$ and variance $\sigma_i^2(x)$, we rely on $v_i(x)$ changing sufficiently slowly relative to $D_i(x)$ so that we can find a subdivision $x_0 \le \cdots \le x_k \le x$ where $v_i(x)$ is approximately constant in each interval. By approximating the mean and variance of $\tau_{x_l\to x_{l+1}}$ with the constant-speed solution, we obtain

$$\mu_i(x) = \int_{x_i(0)}^{x} \frac{dx'}{v_i(x')}, \quad \sigma_i^2(x) = \int_{x_i(0)}^{x} \frac{2D_i(x')}{v_i^3(x')}dx'. \tag{D3}$$

We now compute μ_i and σ_i^2 for the position-dependent speed fluctuations from Ref. [21] of the form

$$v(x) = \overline{v}\,(1 + \delta\cos(\omega x + \phi)) \tag{D4}$$

where x is the distance of the replisome from the origin. We note that because we express $v(x)$ in terms of distance travelled, the speed of the two replisomes when they traverse the same physical location will in general differ. Since integration of $1/v(x)$ and $1/v^3(x)$ is cumbersome, we first compute the following integrals of powers

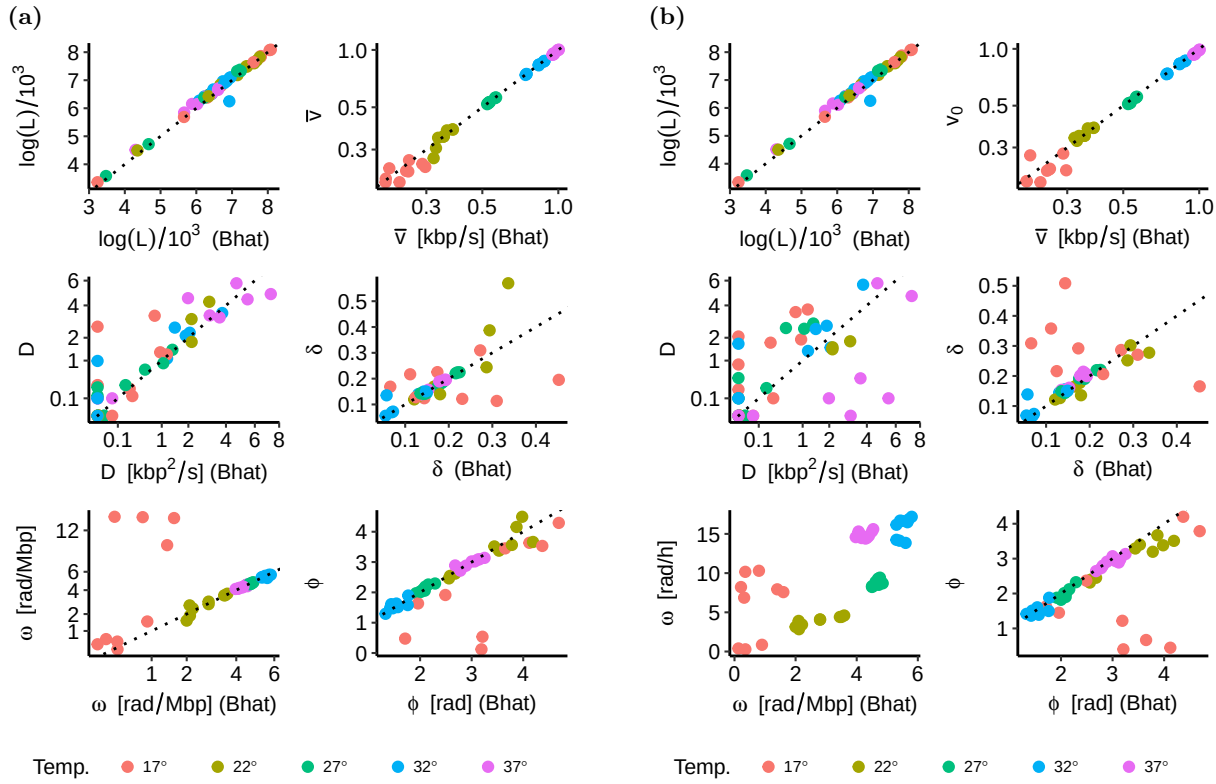

FIG. 6. Comparison of our parameter estimates with those of Ref.[21]. The plots shows data points for all 9 combinations of biological replicate (3 per temperature) and stationary reference sample (3 in total, taken at temperature 17°C, 27°C and 37°C). **(a)**. Position-dependent speed $v(x) = \overline{v}\big(1 + \delta\cos(\omega x + \phi)\big)$. **(b)**. Time-dependent speed $v(\tau) = v_0\big(1 + \delta\cos(\omega\tau + \phi)\big)$.

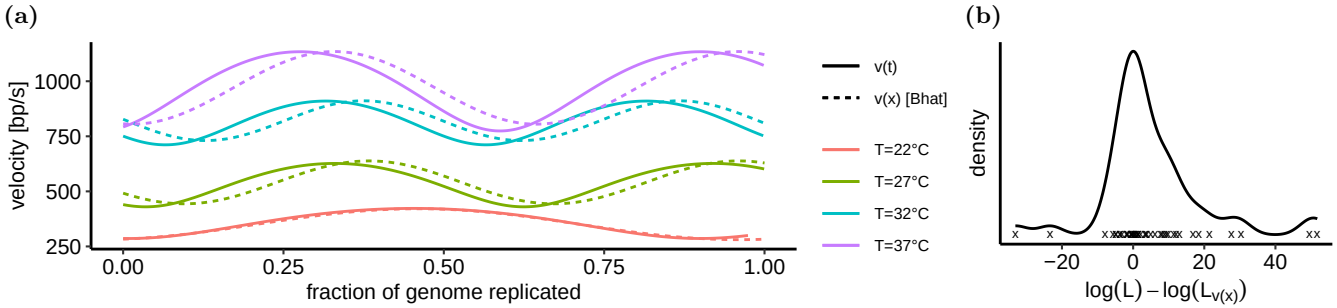

FIG. 7. Time-dependent vs. position-dependent speed oscillations. **(a)**. Time-dependent speed oscillations $v(\tau)$ from figure 3c vs. the position-dependent speed oscillations from Bhat *et al.*[21]. **(b)**. Likelihoods L of the best-fitting time-dependent speed model vs. $L_{v(x)}$ of the best-fitting position-dependent model (computed as described in Appendix E).

of $g(u) = 1 + \delta\cos u$, which are valid for $z \in (-\pi, \pi)$:

$$\hat{G}_1(z) = \int_0^z \frac{du}{g(u)} = \frac{2\tanh^{-1}\left(\frac{(1-\delta)\tan\frac{z}{2}}{\sqrt{1-\delta^2}}\right)}{\sqrt{1-\delta^2}}$$

$$\hat{G}_3(z) = \int_0^z \frac{du}{g^3(u)} = \frac{\left(\delta^2+2\right)\tanh^{-1}\left(\frac{(1-\delta)\tan\frac{z}{2}}{\sqrt{1-\delta^2}}\right)}{\left(1-\delta^2\right)^{5/2}} + \frac{\delta\sin(z)\left(\delta^2 - 3\delta\cos(z) - 4\right)}{2\left(1-\delta^2\right)^2\left(\delta\cos(z)+1\right)^2} \quad \text{(D5)}$$

We also introduce the definite integrals

$$\overline{G}_1 = \int_{-\pi}^{\pi} \frac{du}{g(u)} = \frac{2\pi}{\sqrt{1-\delta^2}},$$

$$\overline{G}_3 = \int_{-\pi}^{\pi} \frac{dz}{g^3(z)} = \frac{\pi\left(\delta^2+2\right)}{\left(1-\delta^2\right)^{5/2}}. \quad \text{(D6)}$$

Writing $[[\cdot]]$ for the nearest integer we obtain

$$G_p(z) = \int_0^z \frac{du}{g^p(u)} = \hat{G}_p(z) + \overline{G}_p\left[\left[\frac{z}{2\pi}\right]\right] \quad \text{(D7)}$$

and in terms of G_p we finally have

$$V_p(x) = \int_0^x \frac{du}{v^p(u)} = \frac{v_0}{\omega}\big(G_p(\omega x + \phi) - G_p(\phi)\big). \quad \text{(D8)}$$

By setting $v_1(x) = v(x)$, $v_2(x) = -v(L-x)$ and $D_1(t) = D_2(t) = D$ in Eq. (D3) and inserting the means and variances into Eq. (D2), we arrive at the replication program

$$f(x,\tau) = 1 - \frac{1}{4}\,\mathrm{erfc}\left(\frac{t - V_1(x)}{2^{\frac{3}{2}} D V_3(x)}\right)\mathrm{erfc}\left(\frac{t - V_1(L-x)}{2^{\frac{3}{2}} D V_3(L-x)}\right). \quad \text{(D9)}$$

Comparison of parameter estimates

Most parameter estimates for the position-dependent speed model obtained using the procedure outlined in Appendix E agree well with the values reported in Ref. [21], see Fig. 6a. Estimates that deviate also exhibit large uncertainties between replicates, see Table II. These large uncertainties can be explained in a similar way as for the time-dependent model presented in Section III.

Parameter estimates also agree well between time- and position-modulated models, see Fig.6b. The oscillation frequency ω is not directly comparable between the two models, but when plotted against replication progress both models yield similar oscillations, see Fig. 7a. When the two models are compared across samples, we find that, for a large majority of samples, the time-dependent speed model yields a slightly higher likelihood than the position-dependent speed model, see Fig. 7b.

Appendix E: Parameter estimation

We compute the position-wise bias-corrected abundances and their uncertainties as

$$a_i = \frac{n_i^{\mathrm{exp}}}{n_i^{\mathrm{stat}}}, \qquad \sigma_i = a_i\sqrt{\frac{1}{n_i^{\mathrm{exp}}} + \frac{1}{n_i^{\mathrm{stat}}}} \quad \text{(E1)}$$

given the experimentally observed read counts at genomic locations $x_1, \ldots, x_n$ in an exponentially growing (n_i^{exp}) and stationary (n_i^{stat}) populations [21]. Since the counts n_i^{exp} and n_i^{stat} are typically large, we approximate the error distributions as Gaussian. Defining the parameter vector $\Theta = \big(v_0, D, \delta, \omega, \phi\big)$, we express the log likelihood by

$$\log L(\Theta) = -\frac{N}{2}\log(2\pi) - \sum_{i=1}^{N}\log(\sigma_i) - \frac{1}{2}\sum_{i=1}^{N}\frac{\big(a_i - \lambda\mathcal{P}(x_i|\Theta)\big)^2}{\sigma_i^2}. \quad \text{(E2)}$$

where $\lambda = \sum_i a_i / \sum_i \mathcal{P}(x_i|\Theta)$ is a scaling factor.

We evaluate $\mathcal{P}(x|\Theta)$ by a numerical integration of Eq.(3). This procedure requires a limited computational effort compared with Ref. [21], that evaluated the likelihood using stochastic simulations. We fit two different models corresponding to two different replication programs with are both instances of the general bacterial program from Eq. (7): (1) the time-dependent speed model in Eq. (9) with the modulating function $h(\tau)$ given in Eq.(13). (2) the position-dependent speed model of Ref. [21] with replication program Eq. (D9), see Appendix D (parameters are $\Theta = \big(\overline{v}, D, \delta, \omega, \phi\big)$ in this case, the parameter $\overline{v}$ takes the place of v_0).

The likelihood surfaces of these models are quite rough, in particular with respect to ω and ϕ. This causes combined gradient-ascent optimization of all five parameters to remain stuck in local maxima, unless the initial estimate is close to the global optimum. On the other hand, exhaustive exploration of the parameter space for all 5 parameters at sufficient resolution is computationally expensive. We thus used the following procedure:

1. We fit the *constant-speed*, *no-diffusion* ($\delta = 0$, $D = 0$) regime by optimizing v_0 while keeping $D = 0$ and $\delta = 0$. Initial value is $v_0 = 500$bp/s.

2. We fit the *oscillatory no-diffusion* ($\delta, \omega, \phi \geq 0$, $D = 0$) regime by optimizing v_0, δ, ω and ϕ while keeping $D = 0$. Initial values lie on a grid formed by the constant-speed, no-diffusion estimate of v_0, $\delta \in \{0, 1/2\}$, 25 logarithmically spaced values for ω from $\pi L/2$ to $32\pi L$ (L being the genome length), and 5 uniformly spaced values for ϕ from 0 to 2π.

3. We fit the *constant-speed diffusive* ($\delta = 0$, $D \geq 0$) regime by optimizing v_0 and D while keeping $\delta = 0$. Initial values lie on a grid formed by the constant-speed, no-diffusion estimate of v_0, and 21 logarithmically spaced values for D from 10 to 10^{10}.

4. We find the initial parameters for the *oscillatory diffusive* ($\delta, \omega, \phi \geq 0$, $D \geq 0$) regime by optimizing v_0, δ, ω and ϕ while fixing D to the *constant-speed diffusive* estimate. Initial values are the same as in (2), except that for v_0 we now use the *constant-speed diffusive* estimate.

5. We fit the *oscillatory diffusive* regime by optimizing v_0, D, δ, ω and ϕ, using the estimates found in (1)-(4) as initial values.

6. We select the parameter regime with the lowest AIC score.

Identifiability

To study the identifiability of parameters v_0, D, δ, ω, ϕ of the bacterial replication program Eq. 9 with $h(t) = v_0\big(1 + \cos(\omega t + \phi)\big)$ we generated artificial DNA abundance distributions and estimated the parameters

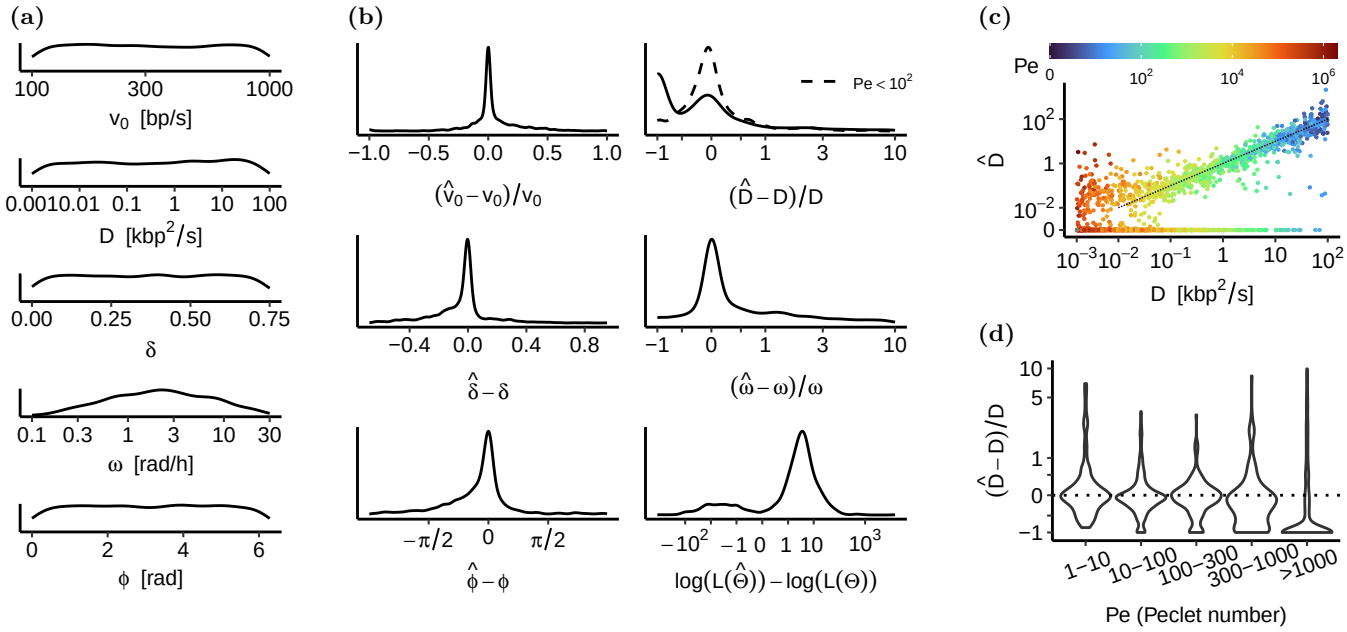

FIG. 8. Parameter recovery. **(a)**. Distributions of true parameters θ. **(b)**. Estimation errors for simulated abundance profiles with $\Lambda = 1\,h^{-1}$ and 40k reads / 10kbp window on average. **(c)**. True value D vs. recovered value $\hat{D}$ for different ranges of the Peclet number $\mathrm{Pe} = Lv_0/4D$. **(d)**. Distribution of the relative error of $\hat{D}$ for different ranges of Pe.

values using the procedure outlined in Appendix E. We randomly selected 2000 sets of parameters (Fig. 8a), and computed their predicted abundance profiles by simulating 10^6 genomes using a fixed growth rate of $\Lambda = 1\mathrm{h}^{-1}$). We then generated Poisson-distributed read counts n_i^{exp} (and similarly for n_i^{stat}, using a flat abundance profile) from these abundance profiles, and ran the estimation procedure described in appendix E. All parameters except D were recovered well (Fig. 8b).

Diffusivities D below a certain value are unlikely to pass the Akaike information criterion and are therefore estimated to be zero. This occurs frequently for parameter sets corresponding to a Peclet number $\mathrm{Pe} = Lv_0/4D$ larger than 300 (Fig. 8c-d).

Appendix F: Eukaryotic DNA replication

In this Appendix, we derive Eq. (21) for the DNA abundance in the case of origins at well-defined locations and with time-homogeneous firing rates, see Eq. (17). The general eukaryotic replication program is given by

$$f(x,\tau) = 1 - e^{-\int\int dx' d\tau' I(x',\tau')\theta\left(\tau-\tau'-\frac{|x-x'|}{v}\right)}, \quad \text{(F1)}$$

see [1, 37]. Equation (F1) is equivalent to Eq. (15), but with an explicit expression for the past light cone. We note that

$$\frac{d}{d\tau}\int\int dx' d\tau' I(x',\tau')\theta\left(\tau-\tau'-\frac{|x-x'|}{v}\right) = \int dx' I\left(x',\tau-\frac{|x-x'|}{v}\right). \quad \text{(F2)}$$

We use Eq. (2) to obtain

$$\mathcal{P}(x) = \int_0^\infty dt\left[\int dx'' I\left(x'',\tau-\frac{|x-x''|}{v}\right)\right]\times \left[e^{-\Lambda\tau-\int\int dx' dt' I(x',\tau')\theta\left(\tau-\tau'-\frac{|x-x'|}{v}\right)}\right]. \quad \text{(F3)}$$

Setting $I(x,\tau) = \sum_{j=1}^K I_j(\tau)\delta(x-x_j)$, i.e. assuming discrete origins $x_1,\ldots,x_K$, yields

$$\mathcal{P}(x) = \int_0^\infty d\tau\left[\sum_j I_j\left(\tau-\frac{|x-x_j|}{v}\right)\right]\times \left[e^{-\Lambda\tau-\sum_j\int_0^{\tau-|x-x_j|/v} d\tau' I_j(\tau')}\right], \quad \text{(F4)}$$

and in the time homogeneous case $I_j(\tau) = I_j^*\theta(\tau)$ we find

$$\mathcal{P}(x) = 1-\Lambda\int_0^\infty d\tau\, e^{-\Lambda\tau-\sum_j I_j^*\left(\tau-\frac{|x-x_j|}{v}\right)\theta\left(\tau-\frac{|x-x_j|}{v}\right)}. \quad \text{(F5)}$$

The integral can be then done by breaking it into different time intervals in which only a certain set of the theta

functions are non-zero. This yields

$$\int_0^\infty d\tau\, e^{-\Lambda\tau-\sum_j I_j^*(\tau-|x-x_j|/v)\theta(\tau-|x-x_j|/v)}$$
$$= \frac{1-e^{-\Lambda\tau_1}}{\Lambda}$$
$$+\sum_{k=1}^{K-1}\frac{e^{-\Lambda\tau_k-\sum_{j=1}^k I_j^*(\tau_k-\tau_j)}-e^{-\Lambda\tau_{k+1}-\sum_{j=1}^k I_j^*(\tau_{k+1}-\tau_j)}}{\Lambda+\sum_{j=1}^k I_j^*}$$
$$+\frac{e^{-\Lambda\tau_K-\sum_{j=1}^K I_j^*(\tau_K-\tau_j)}}{\Lambda+\sum_{j=1}^K I_j^*} \tag{F6}$$

where $\tau_k = \frac{|x-x_k|}{v}$ are the times it takes a replisome formed at the k-th origin to reach position x, and are ordered for g iven x in such a way that $0 \le \tau_1 \le \cdots \le \tau_K$. Therefore, we have

$$\mathcal{P}(x) = e^{-\Lambda\tau_1}$$
$$-\Lambda\sum_{k=1}^{K-1}\frac{e^{-\Lambda\tau_k-\sum_{j=1}^k I_j^*(\tau_k-\tau_j)}-e^{-\Lambda t_{k+1}-\sum_{j=1}^k I_j^*(\tau_{k+1}-\tau_j)}}{\Lambda+\sum_{j=1}^k I_j^*}$$
$$-\Lambda\frac{e^{-\Lambda\tau_K-\sum_{j=1}^K I_j^*(\tau_K-\tau_j)}}{\Lambda+\sum_{j=1}^K I_j^*}. \tag{F7}$$

Equation (21) is obtained from Eq. (F7) by substituting $\mathcal{T}_k$ and W_k as defined in Eqs. (19) and (20) and grouping terms with the same exponent.

Appendix G: Eukaryotic origin inference by simulated annealing

We fit the solution of our eukaryotic model, Eq. (21), to experimentally measured DNA abundances in an exponentially growing asynchronous population of budding yest (*S. cerevisiae* W303)[38]. Also in this case, we correct for sequencing bias using the DNA abundance measured in a stationary population. Our input data $a_1, \ldots, a_N$ are thus the ratios of number of reads in the exponential vs. stationary population found within each 1kb window along the W303 genome (GenBank accession *CM007964.1*). To estimate number and positions of origins and their fork firing rates, we use simulated annealing [46].

Our free parameters are the number of origins K; their positions x_j and intensities I_j^*/v_0, with $j = 1 \ldots K$; and the ratio of growth rate to speed Λ/v_0. We collectively refer to the vector of $2K+1$ free parameters (excluding K itself) as Θ. To prevent over fitting, we use the Akaike information criterion (AIC) as the cost function, defined as

$$\mathrm{AIC}(\Theta) = 2(2\mathrm{K}+1) - 2\log \mathrm{L}(\Theta). \tag{G1}$$

The first term on the right hand side of Eq. (G1) penalizes a large number of origins, while the second term penalizes small log-likelihoods. The quantity $\log L(\Theta)$ is computed as in Eq. (E2). The noise in the data appears larger than expected by Poissonian sampling. We therefore use the empirical noise estimate $\sigma_i = \sigma = \sqrt{\frac{1}{N-1}\sum_i (a_{i+1}-a_i)^2/2} \approx 3.3\cdot 10^{-9}$ instead of the Poissonian error estimates from Eq. (E1). Our estimate is based on the assumption that the systematic contribution to the quantity σ is negligible compared to the noise. We confirmed post-hoc this assumption by replacing the a_i with the fitted model predictions. The resulting value $2\cdot 10^{-10}$ is an order of magnitude smaller; σ therefore indeed mainly represents measurement errors rather than systematic effects.

Our simulated annealing algorithm comprises the following steps:

1. We set the Monte-Carlo temperature to $T_m = 200$ and generate an initial parameter vector Θ. We draw Λ/v_0 uniformly from $[7.7\times 10^{-8}, 2.3\times 10^{-5}]$ bp^{-1} and place two origins randomly on each chromosomes with intensities $I_j^\star/v_0$ drawn uniformly from $[1.1\times 10^{-7}, 1.6\times 10^{-3}]$ bp^{-1}.

2. We generate a parameter proposal Θ' by executing one of the following moves with the indicated probability:

   - *Relocating an origin (prob. 24%).* We choose one among the K origins randomly and relocate it to a uniformly drawn position on the same chromosome. The origin firing rate remains unchanged.
   - *Altering one origin intensity (prob. 24%).* We randomly choose one of the origins and replace its intensity rate $I_j^\star/v_0$ with a value drawn uniformly from $[1.1\times 10^{-7}, 1.6\times 10^{-3}]$ bp^{-1}.
   - *Removing an origin (prob. 24%).* We choose one among the K origins and remove it unless it is the only origin on the chromosome.
   - *Adding an origin (prob. 24%).* We add an origin at a uniformly drawn location on the genome with an intensity uniformly drawn from $[1.1\times 10^{-7}, 1.6\times 10^{-3}]$ bp^{-1}.
   - *Altering the growth rate (prob. 4%).* We draw a new value of Λ/v uniformly from $[7.7\times 10^{-8}, 2.3\times 10^{-5}]$ bp^{-1}.

   The ranges of free parameters are chosen to cover replisome speeds (v) from 10bps^{-1} to 100bps^{-1}, doubling times $\ln 2/\Lambda$ between 50 minutes to 150 minutes, and fork firing delays $1/I_j^*$ between 1 minute and 150 minutes.

3. We accept Θ' with Metropolis probability

$$q = \min\left\{1,\ \exp\left(\frac{\mathrm{AIC}(\Theta)-\mathrm{AIC}(\Theta')}{T_m}\right)\right\}. \tag{G2}$$

4. We reduce the Monte-Carlo temperature by a factor $1 - 1.336 \cdot 10^{-5}$ and reiterate from step (2). The total number of iterations is $2 \cdot 10^6$. With this choice, the final temperature is equal to $T_m = 5 \cdot 10^{-10}$.

After half the number of iterations (10^6), the number of origins is already close to the final value (222 vs. the final number of 234), and repeating the procedure 10 times yield similar final numbers of origins (225 ± 6). We take this as evidence that the algorithm has converged after $2 \cdot 10^6$ iterations.

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

[2] A. Baker, B. Audit, S.-H. Yang, J. Bechhoefer, A. Arneodo, *et al.*, Inferring where and when replication initiates from genome-wide replication timing data, Physical review letters **108**, 268101 (2012).

[3] A. Gispan, M. Carmi, and N. Barkai, Model-based analysis of DNA replication profiles: predicting replication fork velocity and initiation rate by profiling free-cycling cells, Genome Research **27**, 310 (2017).

[4] M. L. Hulke, D. J. Massey, and A. Koren, Genomic methods for measuring DNA replication dynamics, Chromosome Research **28**, 49 (2020).

[5] Z.-Q. Xu and N. E. Dixon, Bacterial replisomes, Current opinion in structural biology **53**, 159 (2018).

[6] D. M. Gilbert, Making sense of eukaryotic DNA replication origins, Science **294**, 96 (2001).

[7] A. Costa and J. F. Diffley, The initiation of eukaryotic DNA replication, Annual Review of Biochemistry **91**, 107 (2022).

[8] S. C.-H. Yang, N. Rhind, and J. Bechhoefer, Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing, Molecular systems biology **6**, 404 (2010).

[9] N. Rhind, DNA replication timing: Biochemical mechanisms and biological significance, BioEssays **44**, 2200097 (2022).

[10] S. Cooper and C. E. Helmstetter, Chromosome replication and the division cycle of Escherichia coli, Journal of Molecular Biology **31**, 519 (1968).

[11] H. Fu, F. Xiao, and S. Jun, Bacterial replication initiation as precision control by protein counting, PRX Life **1**, 013011 (2023).

[12] M. Berger and P. R. ten Wolde, Synchronous replication initiation of multiple origins, PRX Life **1**, 013007 (2023).

[13] 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, Replication dynamics of the yeast genome, science **294**, 115 (2001).

[14] Y. Daigaku, A. Keszthelyi, C. A. Müller, I. Miyabe, T. Brooks, R. Retkute, M. Hubank, C. A. Nieduszynski, and A. M. Carr, A global profile of replicative polymerase usage, Nature structural & molecular biology **22**, 192 (2015).

[15] N. Sueoka and H. Yoshikawa, The chromosome of Bacillus subtilis. I. Theory of marker frequency analysis, Genetics **52**, 747 (1965).

[16] B. M. Wendel, C. T. Courcelle, and J. Courcelle, Completion of DNA replication in Escherichia coli, Proceedings of the National Academy of Sciences **111**, 16454 (2014).

[17] B. M. Wendel, J. M. Cole, C. T. Courcelle, and J. Courcelle, SbcC-SbcD and ExoI process convergent forks to complete chromosome replication, Proceedings of the National Academy of Sciences **115**, 349 (2018).

[18] S. L. Midgley-Smith, J. U. Dimude, T. Taylor, N. M. Forrester, A. L. Upton, R. G. Lloyd, and C. J. Rudolph, Chromosomal over-replication in Escherichia coli recG cells is triggered by replication fork fusion and amplified if replichore symmetry is disturbed, Nucleic acids research **46**, 7701 (2018).

[19] J. U. Dimude, M. Stein, E. E. Andrzejewska, M. S. Khalifa, A. Gajdosova, R. Retkute, O. Skovgaard, and C. J. Rudolph, Origins left, right, and centre: increasing the number of initiation sites in the Escherichia coli chromosome, Genes **9**, 376 (2018).

[20] N. Rhind, S. C.-H. Yang, and J. Bechhoefer, Reconciling stochastic origin firing with defined replication timing, Chromosome Research **18**, 35 (2010).

[21] D. Bhat, S. Hauf, C. Plessy, Y. Yokobayashi, and S. Pigolotti, Speed variations of bacterial replisomes, Elife **11**, e75884 (2022).

[22] D. Huang, A. E. Johnson, B. S. Sim, T. Lo, H. Merrikh, and P. A. Wiggins, The high-resolution in vivo measurement of replication fork velocity and pausing, Nature Communications **14**, 1762 (2022).

[23] D. Huang, T. Lo, H. Merrikh, and P. A. Wiggins, Characterizing stochastic cell-cycle dynamics in exponential growth, Physical Review E **105**, 014420 (2022).

[24] E. Powell, Growth rate and generation time of bacteria, with special reference to continuous culture, Microbiology **15**, 492 (1956).

[25] H. Yoshikawa and N. Sueoka, Sequential replication of Bacillus subtilis chromosome, I. Comparison of marker frequencies in exponential and stationary growth phases, Proceedings of the National Academy of Sciences **49**, 559 (1963).

[26] M. Chandler and R. Pritchard, The effect of gene concentration and relative gene dosage on gene output in Escherichia coli, Molecular and General Genetics MGG **138**, 127 (1975).

[27] F. Jafarpour, C. S. Wright, H. Gudjonson, J. Riebling, E. Dawson, K. Lo, A. Fiebig, S. Crosson, A. R. Dinner, and S. Iyer-Biswas, Bridging the timescales of single-cell and population dynamics, Physical Review X **8**, 021007 (2018).

[28] J. A. Morin, F. J. Cao, J. M. Lázaro, J. R. Arias-Gonzalez, J. M. Valpuesta, J. L. Carrascosa, M. Salas, and B. Ibarra, Active DNA unwinding dynamics during processive DNA replication, Proceedings of the National Academy of Sciences **109**, 8115 (2012).

[29] J. A. Morin, F. J. Cao, J. M. Lázaro, J. R. Arias-Gonzalez, J. M. Valpuesta, J. L. Carrascosa, M. Salas, and B. Ibarra, Mechano-chemical kinetics of DNA replication: identification of the translocation step of a replicative DNA polymerase, Nucleic acids research **43**,

3643 (2015).

[30] S. Jun, H. Zhang, and J. Bechhoefer, Nucleation and growth in one dimension. I. The generalized Kolmogorov-Johnson-Mehl-Avrami model, Physical Review E **71**, 011908 (2005).

[31] S. Jun and J. Bechhoefer, Nucleation and growth in one dimension. II. Application to DNA replication kinetics, Physical Review E **71**, 011909 (2005).

[32] A. Kolmogorov, Statistical theory of nucleation processes., Izu. Akad. Nauk SSSR **3**, 355 (1937).

[33] J. William and R. Mehl, Reaction kinetics in processes of nucleation and growth, Trans. Metall. Soc. AIME **135**, 416 (1939).

[34] M. Avrami, Kinetics of phase change. I General theory, The Journal of chemical physics **7**, 1103 (1939).

[35] K. Sekimoto, Evolution of the domain structure during the nucleation-and-growth process with non-conserved order parameter, Physica A: Statistical Mechanics and its Applications **135**, 328 (1986).

[36] K. Sekimoto, Evolution of the domain structure during the nucleation-and-growth process with non-conserved order parameter, International Journal of Modern Physics B **5**, 1843 (1991).

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

[38] C. A. Müller, M. Hawkins, R. Retkute, S. Malla, R. Wilson, M. J. Blythe, R. Nakato, M. Komata, K. Shirahige, A. P. de Moura, and C. A. Nieduszynski, The dynamics of genome replication using deep sequencing, Nucleic Acids Research **42**, e3 (2013).

[39] A. Shumate and S. L. Salzberg, Liftoff: accurate mapping of gene annotations, Bioinformatics **37**, 1639 (2021).

[40] B. Boldin, O. Diekmann, and J. Metz, Population growth in discrete time: a renewal equation oriented survey, arXiv preprint arXiv:2306.09193 (2023).

[41] A. K. Sinha, C. Possoz, A. Durand, J.-M. Desfontaines, F.-X. Barre, D. R. Leach, and B. Michel, Broken replication forks trigger heritable DNA breaks in the terminus of a circular chromosome, PLoS genetics **14**, e1007256 (2018).

[42] M. R. Evans and S. N. Majumdar, Diffusion with stochastic resetting, Physical review letters **106**, 160601 (2011).

[43] M. R. Evans, S. N. Majumdar, and G. Schehr, Stochastic resetting and applications, Journal of Physics A: Mathematical and Theoretical **53**, 193001 (2020).

[44] M. M. Elshenawy, S. Jergic, Z.-Q. Xu, M. A. Sobhy, M. Takahashi, A. J. Oakley, N. E. Dixon, and S. M. Hamdan, Replisome speed determines the efficiency of the Tus- Ter replication termination barrier, Nature **525**, 394 (2015).

[45] C. A. Nieduszynski, Y. Knox, and A. D. Donaldson, Genome-wide identification of replication origins in yeast by comparative genomics, Genes & Development **20**, 1874 (2006).

[46] D. Bertsimas and J. Tsitsiklis, Simulated Annealing, Statistical Science **8**, 10 (1993).