Inferring Where and When Replication Initiates from Genome-Wide Replication Timing Data

A. Baker,[1] B. Audit,[1] S. C.-H. Yang (楊正炘),[2] J. Bechhoefer,[2] and A. Arneodo[1]

[1]*Université de Lyon, F-69000 Lyon, France, and Laboratoire de Physique, ENS de Lyon, CNRS, F-69007 Lyon, France*
[2]*Department of Physics, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada*

(Received 11 April 2012; published 29 June 2012; publisher error corrected 3 July 2012)

Based on an analogy between DNA replication and one dimensional nucleation-and-growth processes, various attempts to infer the local initiation rate $I(x, t)$ of DNA replication origins from replication timing data have been developed in the framework of phase transition kinetics theories. These works have all used curve-fit strategies to estimate $I(x, t)$ from genome-wide replication timing data. Here, we show how to invert analytically the Kolmogorov-Johnson-Mehl-Avrami model and extract $I(x, t)$ directly. Tests on both simulated and experimental budding-yeast data confirm the location and firing-time distribution of replication origins.

DOI: 10.1103/PhysRevLett.108.268101 PACS numbers: 87.10.−e, 82.60.Nh, 87.14.gk, 87.18.Vf

DNA replication is an essential genomic function responsible for the accurate transmission of genetic information through successive cell generations. Although it is clear that some sites consistently act as replication origins in most eukaryotic cells, the mechanisms that select these sites and the sequences that determine their location remain elusive in many cell types. Furthermore, even less is known about the mechanisms that regulate their firing time [1–4]. Despite recent experimental efforts to map replication origins in higher eukaryotes, the concordance between different studies is generally very low, even when the same technique is employed (e.g., see Ref. [5] for the human genome). Thus, the reliable detection of individual origins is still a very delicate experimental task. This contrasts with the increasing availability of genome-wide replication timing (RT) data for several organisms, ranging from yeast [6] to drosophila [7] to mouse [8] to human [9]. Very recently, genome-wide RT data have been determined in several human cell types [10], providing an unprecedented opportunity to study changes in the replication program that accompany cell differentiation. Given such experimental progress, what kind of information can be extracted about the spatiotemporal replication program? The issue is not trivial, as the RT at a given locus does not necessarily reflect the local initiation properties, because of the confounding effects of passive replication by forks originating from nearby replication origins [11–13]. As a consequence, the observed efficiency of a potential origin depends as much on the context (Is it close or not to other replication origins? What are the firing properties of the neighboring origins?) as on its individual firing properties. The detection and characterization of replication origins from RT data are thus very challenging.

By noticing that the DNA replication program is formally equivalent to a one-dimensional nucleation-and-growth process, Bechhoefer's group has generalized and adapted the Kolmogorov-Johnson-Mehl-Avrami (KJMA) theory of phase transition kinetics [14] to the study of DNA replication kinetics [15]. Assuming that DNA synthesis is bidirectional, that origins fire independently of each other, and that the replication fork velocity v is constant (as recently shown by DNA combing in HeLa cells [16] and by ChIP-chip analysis in yeast [17]), they demonstrated that once the local initiation rate $I(x, t)$ (number of initiations per time per unreplicated length at locus x and time t) is given, most features of the DNA replication program can be analytically predicted, including the observed density of initiation $n(x, t)$ and the RT distribution $P(x, t)$ (probability that locus x is replicated at time t) [13]. Almost all stochastic models of the DNA replication program proposed so far [11–13] assume independent firing of replication origins and are thus special cases of the KJMA model. The spacetime-dependent initiation rate $I(x, t)$ can thus be considered as the basic ingredient of the model. Recently, various groups [11,13,18(c)] have attempted to infer the local initiation properties from RT data in budding yeast, where the position of potential replication origins are well characterized [19]. They all used a fitting strategy that consists in determining the parameters in a previously determined functional form for $I(x, t)$ that best reproduces the RT data. In such approaches, the general form of $I(x, t)$ must be specified in advance, a requirement that can be awkward in the absence of good models of the underlying biology. In addition, such methods will require significant computer resources in order to scale up from yeast genomes (10^7 bases) to mammalian genomes (10^9 bases).

In this Letter, we solve the inverse problem and show that, in the framework of the KJMA model, $I(x, t)$ can be analytically determined from the RT distribution $P(x, t)$ without knowing the functional form of the initiation kinetics, a result that allows analysis of large-scale genomic data using only modest computational resources.

In the generalization of the KJMA model to DNA replication kinetics the unreplicated fraction $s(x, t)$, the fraction of cells where the locus x is not yet replicated at time t, can be directly computed from the initiation rate $I(x, t)$ [13]:

$$s(x,t) = e^{-\int_{V_X(v)} dY I(Y)}, \tag{1}$$

where $V_X(v)$ is the past light cone of the spacetime point $X = (x, t)$ (gray area in Fig. 1). As origins are supposed to fire independently, the probability to observe an initiation at X is $n(x,t) = I(x,t)s(x,t)$. Moreover, since a locus x is unreplicated at time t iff its RT, $t_R(x)$, is greater than t, the RT probability distribution can be derived from $s(x,t) = \mathrm{Prob}[t_R(x) \geq t]$: $P(x,t) = -\partial_t s(x,t)$. Note that because of passive replication [11,13], the observed RT distribution $P(x_i, t)$ at origin i is generally not equal to the intrinsic firing time distribution $\phi(x_i, t)$, defined as the probability that origin i would initiate at time t if there were no passive replication. Equally, the observed efficiency E_i at locus x_i, defined as the fraction of cells where an initiation is observed at x_i, is generally different from the intrinsic efficiency $\mathcal{E}_i = \int_0^{t_{\mathrm{end}}} dt \phi(x_i, t)$ where t_{end} corresponds to the end of the S phase.

An elegant proof of the inverse problem can be established by introducing light-cone coordinates, $x_\pm = x \mp vt$. In these coordinates, the past light cone of X has a simple expression: $V_X(v) = \{Y$ such that $x_+ \leq y_+, y_- \leq x_-\}$ (Fig. 1). From Eq. (1), $\int_{x_+ \leq y_+, y_- \leq x_-} dy_+ dy_- I(y_+, y_-) = -\ln s(x_+, x_-)$. Differentiating with respect to x_+ and x_-, we get $I(x_+, x_-) = \partial_+ \partial_- \ln s(x_+, x_-)$. Back in the original (x,t) coordinates, this last equation becomes

$$I(x,t) = -\frac{v}{2} \Box \ln s(x,t), \tag{2}$$

where $\Box = \frac{1}{v^2}\partial_t^2 - \partial_x^2$ is the d'Alembertian operator. Thus, the KJMA model can be inverted, providing a way to estimate $I(x,t)$ from data on $s(x,t)$ or on $P(x,t) = -\partial_t s(x,t)$. To apply Eq. (2) to finite-resolution, noisy experimental data, we discretize and smooth the numerical derivatives of the $\Box$ operator using the regularization procedure described in Baker's thesis [20]. Equation (2) may be readily generalized to the case where the replication fork velocity v depends on x and t [20].

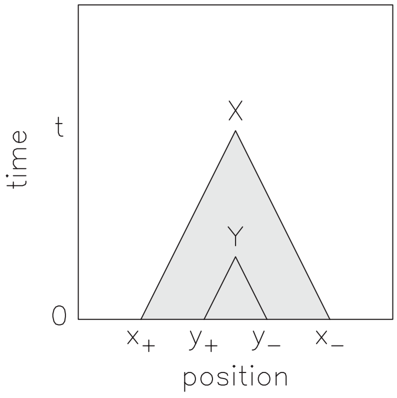

FIG. 1. Light-cone coordinates $x_\pm = x \mp vt$ of spacetime point $X = (x,t)$. Note that $Y = (y,u)$ belongs to the past light cone of X (gray area) iff $x_+ \leq y_+$ and $y_- \leq x_-$.

To illustrate the inversion of the KJMA model on "ideal" RT data, we simulated the unreplicated fraction $s(x,t)$ using the *multiple-initiator model* [13], which fits well the experimental RT data obtained in budding yeast [6]. Figure 2 shows a 260 kbp fragment of yeast chromosome 4 containing 8 potential origins (O_1 to O_8) with intrinsic efficiencies close to 1. (O_7 has the smallest intrinsic efficiency, $\mathcal{E} = 0.96$.) We used the forward KJMA formula, Eq. (1), to generate the theoretical $s(x,t)$ at a fine spatial (2 kbp) and temporal (1 min) resolution. Note that while the spatial resolution corresponds to the resolution of the experimental data fit by Yang, Rhind, and Bechhoefer [13], the temporal resolution is finer than the experimental one (5 min). From the theoretical unreplicated fraction $s(x,t)$ shown in Fig. 2(a), we compute the local initiation rate $I(x,t)$ using the analytical inversion formula Eq. (2). From $I(x,t)$, we then determine the intrinsic firing time distribution $\phi(x,t)$ [Fig. 2(b)], the observed density of initiations $n(x,t)$, the observed efficiency $E(x)$, and the intrinsic efficiency $\mathcal{E}(x)$ (data not shown). We recover in Fig. 2(b) the location of the 8 potential origins of the multiple-initiator model. We can even distinguish the intrinsic firing time distribution of each potential origin; for instance, O_6 tends to fire early while O_7 tends to fire at mid S phase ($t \approx 30$ min). Figure 3 focuses on the origins O_3, O_6, and O_7, where the initiation rate $I(x,t)$ determined by the analytical inversion agrees with the input theoretical initiation rate [Fig. 3(a)], as well as the intrinsic firing time distribution $\phi(x,t)$ [Fig. 3(b)]. We

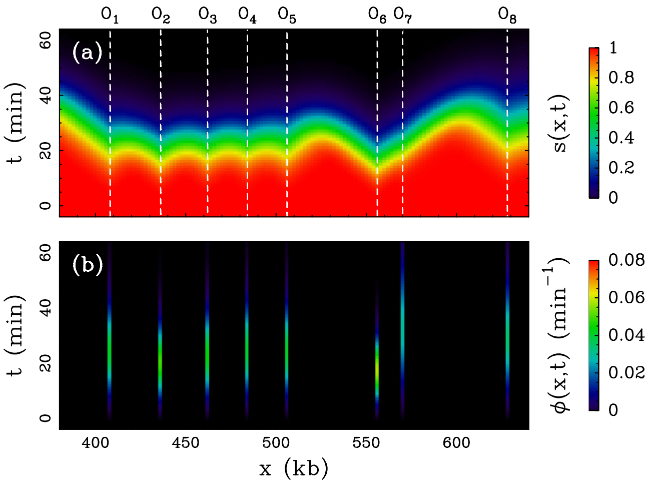

FIG. 2 (color online). On a 260-kbp fragment of yeast chromosome 4, containing 8 potential replication origins (O_1 to O_8), the unreplicated fraction $s(x,t)$ (a) given by the multiple-initiator model [9] was generated by the forward KJMA formula, Eq. (1). The local initiation rate $I(x,t)$ was computed from the unreplicated fraction $s(x,t)$ (a) using the inversion Eq. (2). The intrinsic firing time distribution $\phi(x,t)$ in (b) was then determined according to $\phi(x_i,t) = I(x_i,t)e^{-\int_0^t du I(x_i,u)}$ [13].

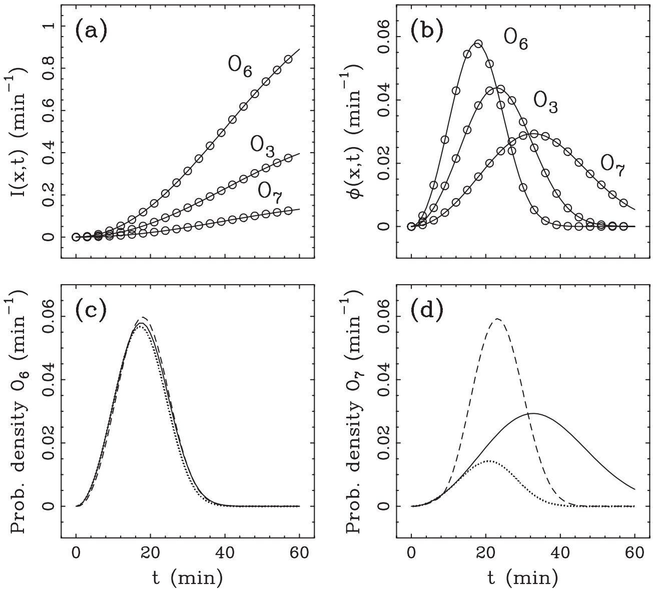

FIG. 3. Comparison of the analytical inversion (circles) with the theoretical solution (solid lines) for potential origins O_3, O_6, and O_7: (a) local initiation rate $I(x, t)$; (b) firing-time distribution $\phi(x, t)$. Quantifying the impact of passive replication by comparing the intrinsic firing time distribution $\phi(x, t)$ (solid line), the RT distribution $P(x, t)$ (dashed line), and the observed density of initiation $n(x, t)$ (dotted line): (c) potential origin O_6; (d) potential origin O_7.

notice in Fig. 2(a) that the origin O_7, detected by the numerical inversion, does not correspond to a local minima of the unreplicated fraction. About one origin in three in the multiple-initiator model is not associated with a local minimum in the unreplicated fraction data [13]. It is sometimes assumed that origins whose positions are well defined correspond to local minima in the mean RT or the unreplicated fractions [21]. Such methods would fail to detect the well-positioned origin O_7.

Passive replication can strongly affect both the replication kinetics at a locus and the observed efficiencies of replication origins and can lead to misinterpretation of RT data [11–13]. For a potential origin that is rarely passively replicated, we expect the RT to be equal to the intrinsic firing time of the origin. That is, its RT distribution $P(x, t)$ should be similar to the intrinsic firing time distribution $\phi(x, t)$. Since, in such cases, the firing time corresponds to an observed initiation event, we also expect the observed density of initiations $n(x, t)$ to be similar to the intrinsic firing time distribution $\phi(x, t)$. Indeed, in Fig. 3(c), we see that the early-firing origin O_6, which Fig. 2 shows is unlikely to be passively replicated, has $P(x, t) \sim n(x, t) \sim \phi(x, t)$. These approximations do not hold when the potential origin is passively replicated. For instance, the RT distribution of potential origin O_7, which is often passively replicated by a fork originating from O_6 (Fig. 2), clearly differs from its intrinsic firing time distribution [Fig. 3(d)]. Indeed, the RT distribution of O_7 is close to that of O_6, delayed by the time (7 min) necessary for a fork to propagate from O_6 to O_7 ($x_7 - x_6 = 14$ kbp and $v = 2$ kbp/min). At the onset of S phase ($t < 16$ min), the origin O_7 is unlikely to be passively replicated, since only a few forks coming from O_6 reach O_7 in time, and the observed density of initiations at O_7 is very similar to its intrinsic firing time distribution [Fig. 3(d)]. At later times, the observed density of initiations at O_7 is strongly reduced, as O_6 becomes more likely to passively replicate O_7. Even though the origin O_7 has an intrinsically high probability of firing for $t > 30$ min, we almost never observe initiations at those times [Fig. 3(d)]. Because of the context (the early-firing origin O_6 located nearby), the observed density of initiations and the RT at O_7 are strongly affected by the passive replication of O_7.

Figure 4 reports results from the first application of the analytical inversion formula Eq. (2) to the RT microarray data obtained by McCune *et al.* for budding yeast [6]. As pointed out in Ref. [13], these data suffer from severe

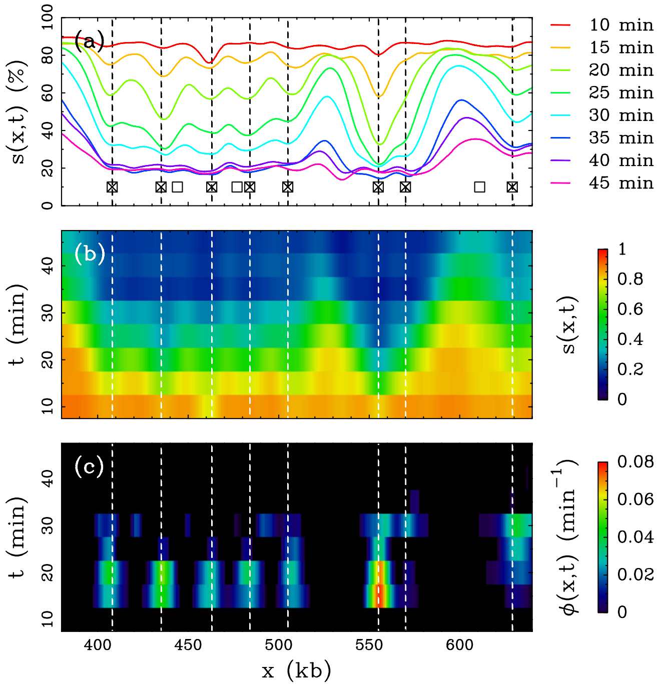

FIG. 4 (color online). (a) Experimental unreplicated fraction profiles $s(x, t)$ obtained at different S phase times (from $t = 10$ min to $t = 45$ min), along the same fragment of yeast chromosome 4 as in Fig. 2 (the data were retrieved from Ref. [6]). The (☒) correspond to potential origins (O_1 to O_8) whereas the (□) correspond to false origins or to origins that do not contribute enough to the replicated fraction (see Ref. [13] for the detailed criteria for elimination from the 732 origins recorded in the OriDB database [19]). (b) Two-dimensional spatiotemporal representation of $s(x, t)$. (c) Intrinsic firing time distribution $\phi(x, t)$ obtained from the experimental $s(x, t)$ by solving Eq. (2) [see Fig. 2(b)]. Note that before applying Eq. (2), the experimental unreplicated fraction data were modified, as explained in the text, to render them compliant with the KJMA kinetics.

artifacts and limitations, including finite temporal ($\Delta_t = 5$ min) and spatial ($\Delta_x = 2$ kbp) resolution, uncertainty in S phase duration, starting-time asynchrony of cell cycles, and limited range of replication fraction (0–100%), probably because of contamination or imperfect signal normalization. Along the lines of the strategy used in Ref. [13], prior to applying Eq. (2) to the experimental $s(x, t)$, we have modified it to make it compliant with the KJMA kinetics. Three main steps were taken to clean up the RT data: (1) Causality requires that $s(x, t)$ decreases with time t. We thus changed iteratively the unreplicated fraction according to $s(x, t + \Delta_t) \leftarrow \min[s(x, t + \Delta_t), s(x, t)]$. (2) If replication forks propagate at velocity v, then for each spacetime point X and for every Y in the past light cone of X, we have $s(X) \leq s(Y)$. To satisfy this requirement, it is sufficient to change iteratively the unreplicated fraction according to $s(x, t + \Delta_t) \leftarrow \min[s(x, t + \Delta_t), \min_{y\in[x-\Delta_x,x+\Delta_x]} s(y, t)]$, where $\Delta_x = v\Delta_t$ with $v = 2$ kbp/ min. (3) The independent firing of replication origins implies that $I(X)$ is positive for any spacetime point X. This requirement is equivalent to $s(x, t + \Delta_t) \leq \frac{s(x+\Delta_x,t)s(x-\Delta_x,t)}{s(x,t-\Delta_t)}$. We therefore changed iteratively the unreplicated fraction according to $s(x, t + \Delta_t) \leftarrow \min[s(x, t + \Delta_t), \frac{s(x+\Delta_x,t)s(x-\Delta_x,t)}{s(x,t-\Delta_t)}]$.

The experimental unreplicated fraction $s(x, t)$ for the 260 kbp yeast fragment previously investigated (Fig. 2) is shown in Figs. 4(a) and 4(b). We modified the unreplicated fraction according to the three steps described above and, on the resulting unreplicated fraction, we applied Eq. (2) to obtain the local initiation rate $I(x, t)$, from which we deduced the intrinsic firing time distribution $\phi(x, t)$ shown in Fig. 4(c). When comparing with the $\phi(x, t)$ obtained numerically in Fig. 2(b), we confirm that the inversion works well qualitatively. Among the 11 origins of replication recorded along this yeast fragment in the OriDB database [19], we recover the locations of the 8 potential origins O_1 to O_8 identified by Yang *et al.* after eliminating false (and/or too weak) origins [13]. Furthermore, the firing time distribution [Fig. 4(c)] agrees qualitatively with the firing time distribution of the multiple-initiator model [Fig. 2(b)]; for instance O_6 is a very efficient, early-firing origin while its neighbor O_7 is much less efficient because of passive replication and fires at $t \approx 30$ min. Let us point out that we have tried a number of other, equally "arbitrary" methods of preprocessing the McCune *et al.* RT microarray data [6] and found much the same results as those reported in Fig. 4 (data not shown).

In summary, we have shown how to extract the local initiation rate $I(x, t)$—where and when replication initiates—from RT data without preexisting models. Our assumptions—causality, constant replication fork velocity, and independent firing of replication origins—are modest. In practice, this inverse strategy depends on high-quality RT data (with good spatial and temporal resolution, high signal-to-noise ratio, etc.) as well as efficient regularization schemes to solve Eq. (2). Future applications to RT data from higher eukaryotic organisms are promising. For the human genome, application of the present theoretical results should enrich the set of 1000–2000 origins recently detected as local RT minima bordering megabase-sized RT domains in various human cell lines [22]. As originally identified in the germ line from asymmetry in the composition profile of human chromosomes [23], these "master" origins are specified by a region of open chromatin in an otherwise heterochromatin environment [22,24] and are likely to play a key role in the regulation of the replication spatiotemporal program by chromatin tertiary structure [22,25]. Detecting "secondary" replication origins would help test recently proposed scenarios, including a chromatin-mediated succession of independent secondary activations [16] and a dominolike cascade of secondary activations induced in front of the propagating replication fork [12,16]. Further analyses of RT data are underway to explore these issues.

We thank Y. d'Aubenton-Carafa, A. Goldar, O. Hyrien, C. Thermes, and C. Vaillant for interesting discussions. This work was supported by the ANR (REFOPOL, ANR 10 BLAN 1615) and by NSERC (Canada).

[1] D. M. Gilbert, Science **294**, 96 (2001).
[2] M. I. Aladjem, Nat. Rev. Genet. **8**, 588 (2007).
[3] M. Méchali, Nat. Rev. Mol. Cell Biol. **11**, 728 (2010).
[4] *DNA Replication and Human Disease*, edited by M. L. DePamphilis (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 2006).
[5] J.-C. Cadoret, F. Meisch,V. Hassan-Zadeh, I. Luyten, C. Guillet, L. Duret, H. Quesneville, and M.-N. Prioleau, Proc. Natl. Acad. Sci. U.S.A. **105**, 15 837 (2008); N. Karnani, C. M. Taylor, and A. Dutta, Methods Mol. Biol. **556**, 191 (2009); J. L. Hamlin, L. D. Mesner, and P. A. Dijkwel, Chromosome Res. **18**, 45 (2010).
[6] H. J. McCune, L. S. Danielson, G. M. Alvino, D. Collingwood, J. J. Delrow, W. L. Fangman, B. J. Brewer, and M. K. Raghuraman, Genetics **180**, 1833 (2008).
[7] D. Schübeler, D. Scalzo, C. Kooperberg, B. van Steensel, J. Delrow, and M. Groudine, Nat. Genet. **32**, 438 (2002).
[8] I. Hiratani *et al.*, Genome Res. **20**, 155 (2010).
[9] K. Woodfine, D. M. Beare, K. Ichimura, S. Debernardi, A. J. Mungall, H. Fiegler, V. P. Collins, N. P. Carter, and I. Dunham, Cell Cycle **4**, 172 (2005).
[10] R. Desprat, D. Thierry-Mieg, N. Lailler, J. Lajugie, C. Schildkraut, J. Thierry-Mieg, and E. E. Bouhassira, Genome Res. **19**, 2288 (2009); C. L. Chen *et al.*, Genome Res. **20**, 447 (2010); R. S. Hansen, S. Thomas, R. Sandstrom, T. K. Canfield, R. E. Thurman, M. Weaver, M. O. Dorschner, S. M. Gartler, and J. A. Stamatoyannopoulos, Proc. Natl. Acad. Sci. U.S.A. **107**, 139 (2010); T. Ryba *et al.*, Genome Res. **20**, 761 (2010); E. Yaffe, S. Farkash-Amar, A. Polten, Z. Yakhini, A. Tanay, and I. Simon, PLoS Genet. **6**, e1001011 (2010).

[11] A. P. S. de Moura, R. Retkute, M. Hawkins, and C. A. Nieduszynski, Nucleic Acids Res. **38**, 5623 (2010); R. Retkute, C. A. Nieduszynski, and A. P. S. de Moura, Phys. Rev. Lett. **107**, 068103 (2011).

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

[13] S. C.-H. Yang, N. Rhind, and J. Bechhoefer, Mol. Syst. Biol. **6**, 404 (2010).

[14] J. Christian, *The Theory of Phase Transformations in Metals and Alloys. Part I: Equilibrium and General Kinectics Theory* (Pergamon, Oxford, 2002).

[15] S. Jun, H. Zhang, and J. Bechhoefer, Phys. Rev. E **71**, 011908 (2005); S. Jun and J. Bechhoefer, Phys. Rev. E **71**, 011909 (2005); H. Zhang and J. Bechhoefer, Phys. Rev. E **73**, 051903 (2006); J. Bechhoefer and B. Marshall, Phys. Rev. Lett. **98**, 098105 (2007); M. G. Gauthier and J. Bechhoefer, Phys. Rev. Lett. **102**, 158104 (2009).

[16] G. Guilbaud *et al.*, PLoS Comput. Biol. **7**, e1002322 (2011).

[17] M. D. Sekedat, D. Fenyö, R. S. Rogers, A. Tackett, J. D. Aitchison, and B. T. Chait, Mol. Syst. Biol. **6**, 353 (2010).

[18] (a) J. Lygeros, K. Koutroumpas, S. Dimopoulos, I. Legouras, P. Kouretas, C. Heichinger, P. Nurse, and Z. Lygerou, Proc. Natl. Acad. Sci. U.S.A. **105**, 12295 (2008); (b) J. J. Blow and X. Q. Ge, EMBO Rep. **10**, 406 (2009); (c) H. Luo, J. Li, M. Eshaghi, J. Liu, and R. M. Karuturi, BMC Bioinf. **11**, 247 (2010).

[19] C. A. Nieduszynski, S.-I. Hiraga, P. Ak, C. J. Benham, and A. D. Donaldson, Nucleic Acids Res. **35**, D40 (2007).

[20] A. Baker, Ph.D. thesis, University of Lyon, France, 2011.

[21] M. K. Raghuraman *et al.*, Science **294**, 115 (2001).

[22] A. Baker *et al.*, PLoS Comput. Biol. **8**, e1002443 (2012).

[23] E.-B. Brodie, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, C. Thermes, and A. Arneodo, Phys. Rev. Lett. **94**, 248103 (2005); M. Touchon, S. Nicolay, B. Audit, E.-B. Brodie of Brodie, Y. d'Aubenton-Carafa, A. Arneodo, and C. Thermes, Proc. Natl. Acad. Sci. U.S.A. **102**, 9836 (2005); M. Huvet, S. Nicolay, M. Touchon, B. Audit, Y. d'Aubenton-Carafa, A. Arneodo, and C. Thermes, Genome Res. **17**, 1278 (2007); B. Audit, S. Nicolay, M. Huvet, M. Touchon, Y. d'Aubenton-Carafa, C. Thermes, and A. Arneodo, Phys. Rev. Lett. **99**, 248102 (2007).

[24] B. Audit, L. Zaghloul, C. Vaillant, G. Chevereau, Y. d'Aubenton-Carafa, C. Thermes, and A. Arneodo, Nucleic Acids Res. **37**, 6064 (2009).

[25] P. St-Jean, C. Vaillant, B. Audit, and A. Arneodo, Phys. Rev. E **77**, 061923 (2008); A. Arneodo, C. Vaillant, B. Audit, F. Argoul, Y. d'Aubenton-Carafa, and C. Thermes, Phys. Rep. **498**, 45 (2011).