

Computational methods to study kinetics of DNA replication

Scott Cheng-Hsin Yang, Michel G. Gauthier, and John Bechhoefer

Dept. of Physics, Simon Fraser University, Burnaby, BC, V5A 1S6, Canada

Abstract

New technologies such as DNA combing have led to the availability of large quantities of data that describe the state of DNA while undergoing replication in S phase. In this chapter, we describe methods used to extract various parameters of replication — fork velocity, origin initiation rate, fork density, numbers of potential and utilized origins — from such data. We first present a version of the technique that applies to “ideal” data. We then show how to deal with a number of real-world complications, such as the asynchrony of starting times of a population of cells, the finite length of fragments used in the analysis, and the finite amount of DNA in a chromosome.

Key words: DNA replication, replication fork velocity, origin initiation

1 Introduction

New techniques, DNA combing in particular [1], have led to the possibility of obtaining large quantities of data on the progress of DNA replication. Over the past few years, such experiments have been conducted on a number of different organisms, including *Xenopus* cell-free embryo extracts [2], the Epstein-Barr virus [3], budding yeast [4], Chinese Hamster cells [5], fission yeast [6], and human cancer cells [7]. These experiments have led to important insights into the mechanisms of DNA replication, such as the role that the origin-initiation rate plays in the successful completion of replication [8,9,10], the role of chromatin looping [11], and a much more detailed appreciation of the complicated interplay between origin initiation and fork progression [12].

Molecular combing experiments generate large quantities of data that typically take the form of images of fragments of DNA, with various domains labeled. In the simplest example, described in this chapter, the labeled regions correlate with replicated or non-replicated domains, allowing one to have, in effect, a

“snapshot” of the state of the DNA at some time point during replication. The goal of the analytical methods presented here is to extract from such data parameters that are relevant to DNA replication. These parameters include the replication fork velocity, origin-initiation rate, fork density, and numbers of potential and initiated origins. The parameters are described by fitting statistics of the combing data — for example, the average size of replicated or non-replicated domains — to a kinetic model of DNA replication. Here, a “kinetic model” is one that seeks to *describe* the progress of replication in a way that is independent of the underlying biological mechanisms. For example, a key parameter in such models is the time-dependent rate of initiation of origins, $I(t)$. This is a kinetic parameter in the sense that one can describe the number of origin initiations independently of the mechanism that determines these numbers. In the work on *Xenopus* cell-free embryo extracts, $I(t)$ was found using a kinetic model to increase throughout S phase. The inferred measurements of $I(t)$ then stimulated several hypotheses about possible underlying biological mechanisms to account for this increase [13,14,15]. Although one might view the relative lack of biological detail in kinetic models as a limitation, it can also be an advantage, in that one can separate the description of the progress of DNA replication from any explanation of mechanisms.

Below, we describe how to use kinetic models to extract from experimental data various replication parameters such as the velocity of replication forks, the numbers of potential and activated origins, and the rate of initiation of origins. In the Materials Section, we outline the data requirements for the analysis. In the Methods Section, we describe the structure of the basic model and give an analysis that is appropriate to “ideal” data. Then, in the Notes Section, we describe various complications that are likely to be present in typical data and give methods for adapting the basic analysis to deal with the complications.

2 Materials

The input data for kinetic modeling have been data derived from molecular combing experiments. These data are in the form of fragments of fluorescently labeled DNA, imaged by epifluorescence microscopy, and recorded digitally. In the following, we consider the simplest example of a combing experiment, where at a given time point during the replication cycle (i.e., at time t after the start of S phase), a nucleotide analog such as bromodeoxyuridine (BrdU, an analog of thymidine) is introduced and incorporated into the replicating DNA. After replication is completed, the DNA is extracted, combed onto a substrate, and the BrdU is labeled using an anti-BrdU antibody labeled with a fluorescent dye such as the fluorescein-based FITC. In addition, the entire DNA fragment is labeled with a non-specific label with an anti-guanosine

antibody attached to a different-color fluorophore, in order to visualize the entire fragment. Both labels are then imaged, allowing one to infer a kind of snapshot of the replication state of the DNA at the time that the BrdU was added (Fig. 1). The experiment is then repeated for different time points, giving information about the replication state as the cell progresses through S phase. Further details on molecular combing of DNA for replication studies are given in other chapters in this volume.

The images of combed fragments are analyzed, either manually via an image-processing program or by specialized software such as that available from Genomic Vision (www.genomicvision.com). For the former strategy, the open-source ImageJ (rsb.info.nih.gov/ij) is a common choice. One uses a measuring tool to determine the lengths of labeled domains and DNA fragments, using one's eye to determine the domain boundaries. The resulting data set has one record per analyzed fragment. Figure 1 shows a schematic of a typical fragment. The thick black lines represent domains of replicated DNA (“eyes”); the thin ones domains that had not yet replicated at the time the labels were introduced to the sample (“holes”). A final quantity of interest is the “eye-to-eye” distance, defined to be the distance between the centers of two neighboring eyes.

The initial task, then, is to compile a list, for each fragment, of data obtained via image analysis. This may be done either with a spreadsheet program such as Excel (Microsoft, Inc.), or an open-source equivalent such as Calc (www.openoffice.org). Alternatively, a more-sophisticated scientific data-analysis tool such as Igor-Pro (WaveMetrics, Inc.; used in our own work) or Matlab (The MathWorks, Inc.) may be used. The latter programs have the advantage of being able to carry out Monte Carlo simulations of DNA replication, and one can use the resulting simulation data as substitutes for analytical functions when fitting to experimental data. This can be important in that deriving analytical expressions for realistic models may be hard. However, at least for the simpler versions of the analysis discussed here, a spreadsheet will suffice.

Table 1, below, illustrates a typical sample format for data corresponding to the image in Fig. 1. As we discuss below, the quality of the raw data influences both the reliability of the inferred replication parameters and the computational effort required to extract those parameters. The most important considerations are as follows:

- (1) The cell populations should be well-synchronized. In other words, if data are taken from a population of cells, those cells should all have started replication at approximately identical times. In particular, the standard deviation of the starting times should be much less than the duration of S phase. (See Note 4.1.)

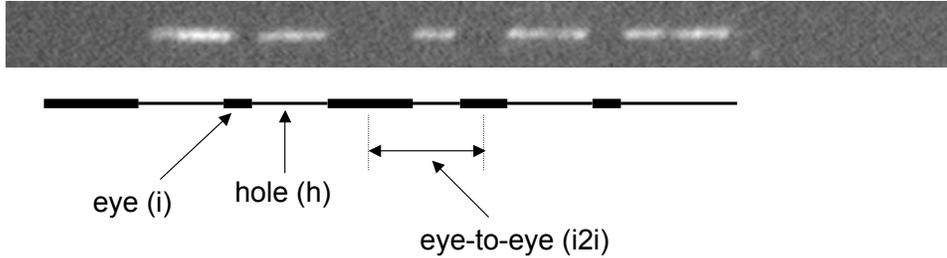


Fig. 1. Top: Epifluorescence image of a combed fragment of DNA labeled to show non-replicated areas. Non-replicated segments are visualized using anti-BrdU antibodies. The length and continuity of the DNA fragment is determined by labeling with anti-guanosine antibodies (image not shown). Bottom: Schematic diagram corresponding to the labeled fragment of DNA, resulting from a molecular-combing experiment. Eye, hole, and eye-to-eye domain sizes are indicated. Combing image courtesy John Herrick, Genomic Vision.

- (2) The combed fragments of DNA should be as large as possible. As we discuss below in Note 4.2, the finite length of combed DNA fragments can bias the measurement of average domain sizes downwards. Since we use measurements of average eye and hole sizes in the determination of origin initiation rates, etc., their estimates can also be biased. The important measure of fragment length is not an absolute length but the average number of domains (eyes, holes) per fragment, $N_{domains}$. Near the beginning of S phase, the eyes are small and holes are large, and the reverse is true at the end of S phase. In both cases, it is clear that a typical fragment will have few domains. Thus, $N_{domains}$ will be largest in the middle of S phase. If $N_{domains} > 10$, then finite-size effects are small.
- (3) Good optical resolution and good labeling efficiency are also important. Here, the goal is to minimize the number of mistakes made in the domain assignment. These can arise when a very small domain (say an eye) is not well-resolved, leading one to confuse a hole-eye-hole sequence with a single larger hole. The reverse scenario is that non-specific labeling causes one to misinterpret a large hole with a false hole-eye-hole sequence. A reasonable criterion is to limit such mis-assignments to no more than 1% of the total amount of data gathered.
- (4) Finally, the total amount of data is also important. As a rule of thumb, one should have data from DNA fragments whose total length exceeds that of the original genome. However, multiple coverage is better.

| | |
|--------------------|------|
| fragment label | 13 |
| fragment length | 38 |
| number of domains | 4 |
| length of domain | 0 |
| “ ” | 18 |
| “ ” | 15 |
| “ ” | 5 |
| end of record code | 9999 |

Table 1

Sample data obtained from analysis of an image of a combed DNA fragment. The type of each domain alternates between hole and eye. By convention, the first domain is a hole. Since the above fragment begins with an edge eye domain, there is a fictitious zero-length entry and hence only 3 actual domains. All lengths are given in kilobases (kb). The 9999 entry is a redundant marker to aid in the reading of the data file.

3 Methods

3.1 Kinetic Modeling Approach

In replication experiments, the quantities of direct interest include the replication fork velocity v , the rate of initiation of origins I , the fork density n_f – all of which can depend on time during S phase and location along the genome. In principle, if one were able to image the replication process dynamically, such quantities could be extracted in a straightforward manner. But given that data from actual experiments have been limited to static snapshots of the replication state of DNA fragments, a more sophisticated approach is needed. To understand why a statistical approach is necessary, consider again the image and sketch of Fig. 1. There, several replicated domains (eyes) are indicated. Can we conclude that there was a single origin at the center of each eye? Unfortunately, no: while it is possible that an eye is the result of a single initiation event, it could also result from two or more initiations that subsequently merged. Thus, it is not straightforward to use the number of eye domains at different times to infer the rate of origin initiation. One might try an *ad hoc* approach, where eyes below a certain small size are deemed to be the result of a single initiation. One study, for example, used data from eyes with size between 3 and 8 kb [2]. (The lower limit comes from the need to distinguish between a domain and a non-specifically bound fluorophore.) While such an approach can give some information on the time-dependent initiation

rate $I(t)$, one is throwing away most of the data and thus increasing statistical errors. In addition, biases will arise if the small domains examined actually do correspond to two or more initiation sites or if domains larger than the cutoff have just a single origin.

The kinetic-modeling approach presented here skirts these difficulties. Because the model is statistical, it can incorporate all the acquired data. In effect, there is no need to decide whether a given domain has one or more origins. The quantities of interest become statistics of domain sizes – for example, the average eye, hole, and eye-to-eye sizes. (Higher-moment statistics such as the standard deviation can give more information but have not so far been exploited, as their accurate estimation would require more data than have typically been available.)

The models that we use have been adapted from earlier work dating from the 1930s on crystallization kinetics [16,17,18]. We emphasize that the analogy is formal and mathematical – not physical or biological. Rather, all that is used are three fundamental aspects of DNA replication: initiation at multiple sites along the genome, outward progression of replication forks, and coalescence of forks that meet. (For the latter, no detailed mechanism is necessary. The assumption essentially is equivalent to the observation that DNA is replicated only once per cell cycle [19].)

In this section, we assume “good” data in that the various desiderata in the Methods Section, above, have been met. In fact, they typically have not been met in experiments conducted to date, and in the Notes Section, below, we explain how to deal with various non-ideal situations encountered in practice. The analysis given below, however, is useful both for its relative simplicity and because an experiment giving data that are good enough to be analyzed as done below could in principle be carried out. Note that we only cite the main results; for derivations, see [20] and references therein.

Two principle results are an expression for the domain density $n_d = N_d/L$, with N_d the number of domains in a genome of length L ,

$$n_d(t) = g(t)e^{-2v \int_0^t g(t')dt'} , \quad (1)$$

and an expression for the overall fraction of the genome that has replicated at time t ,

$$f(t) = 1 - e^{-2v \int_0^t g(t')dt'} , \quad (2)$$

where $g(t) = \int_0^t I(t')dt'$, with $I(t)$ the initiation rate. A glossary of technical symbols is given in Table 2. The domain density is a bell-shaped curve peaking

| symbol | definition |
|-------------------------|---|
| f | replication fraction ($0 < f < 1$) |
| I | initiations / length of unreplicated DNA / time |
| $g(t)$ | integral of I from time 0 to time t |
| v | replication fork velocity (kb/min) |
| $N_{domains}$ | number of domains / DNA fragment of length L |
| $n_{domains}$ | average number of domains / length of DNA |
| n_o | number of initiated origins / length of DNA |
| $\bar{\ell}_i$ | average length of replicated domains (“eyes”) |
| $\bar{\ell}_h$ | average length of non-replicated domains (“holes”) |
| $\bar{\ell}_{i2i}$ | average distance between centers of adjacent replicated domains (“eye-to-eye”) |
| $L_{interior}$ | total length of interior domains |
| L_{edge} | total length of edge domains |
| $L_{oversized}$ | total length of oversized domains |
| $\bar{\ell}_{interior}$ | biased domain-length estimator using only interior domains |
| $\bar{\ell}_{unbiased}$ | unbiased domain-length estimator from interior, edge, and oversized domains |
| t | time elapsed since start of replication |
| τ | laboratory time |
| τ_i | times at which replication data are collected |
| $\phi(\tau)$ | distribution of starting times of DNA replication for different cells |
| $\rho(f, \tau_i)$ | distribution of replication-fraction values of DNA fragments collected at time τ_i |
| $\rho_{end}(t)$ | distribution of replication times for a finite genome |
| t^* | typical time to replicate completely a genome (mode of end-time distribution) |
| β | width (in time) of end-time distribution (\propto standard deviation) |

Table 2

Glossary of technical symbols.

near the middle of S phase, while $f(t)$ is sigmoidal, going from 0 to 1. It is easy to see why the domain density is bell-shaped: at the beginning of S phase, there is a small number of widely separated replicated domains (eyes) and hence a low number of domains/length. At the end of S phase, there are a few widely separated non-replicated domains (holes) and, again, a low domain density. (There is always an equal number of eyes and holes.) In the middle of S phase, there is a relatively large number of medium sized eyes and holes.

As a simple example, if origins initiate at a constant rate, so that $I(t) = I_0$,

then $g(t) = I_0 t$ and Eqs. 1 and 2 imply

$$n_d(t) = I_0 t e^{-I_0 v t^2} ; \quad f(t) = 1 - e^{-I_0 v t^2} . \quad (3)$$

Note that there is a typical time t_0 associated with the replication process that sets the scale for replication times. In this example, it is $t_0 = 1/\sqrt{I_0 v}$, and to progress from 10 to 90% replication requires a time of $\approx 1.2t_0$. (The exact numerical factor depends on the precise form of $I(t)$ as well as the parameters v and I_0 .)

While chromatin and its associated DNA are embedded in a three-dimensional space, they are one-dimensional objects, and that fact imposes constraints on the domain topology. As a result, hole and eye domains must alternate and, as a consequence, one can show that

$$\bar{\ell}_{i2i}(t) = \bar{\ell}_i(t) + \bar{\ell}_h(t) = \frac{1}{n_d(t)} \quad (4)$$

$$f(t) = \frac{\bar{\ell}_i(t)}{\bar{\ell}_i(t) + \bar{\ell}_h(t)} . \quad (5)$$

In Eqs. 4 and 5, the overbars denote averages taken over the set of domain sizes (eye or hole). The first part of Eq. 4 states that the average distance between the centers of two neighboring eyes equals the average size of an eye plus the average size of a hole. The second part, equivalent to $n_d = 1/\bar{\ell}_{i2i}$, states that the domain density is the reciprocal of the average eye-to-eye distance. Equation 5 states that the fraction replicated is the average eye size divided by the average eye plus average hole sizes. Thus, there are only two independent quantities among $f(t)$, $\bar{\ell}_i(t)$, $\bar{\ell}_h(t)$, $\bar{\ell}_{i2i}(t)$, $n_d(t)$. Then Eqs. 1, 2, 4, and 5 together imply

$$\bar{\ell}_i(t) = \frac{1}{g(t)} \left[e^{2v \int_0^t g(t') dt'} - 1 \right] \quad (6)$$

$$\bar{\ell}_h(t) = \frac{1}{g(t)} \quad (7)$$

$$\bar{\ell}_{i2i}(t) = \frac{1}{g(t)} e^{2v \int_0^t g(t') dt'} . \quad (8)$$

Finally, the number of initiated origins per unit length along the genome, n_o , can be written as

$$n_o = \int_0^\infty I(t) [1 - f(t)] dt . \quad (9)$$

3.2 Extraction of Replication Parameters using the Kinetic Approach

In the Materials section, we outlined the collection of data under “ideal” circumstances — many long fragments of DNA with numerous domains, highly efficient and specific labeling, and all taken from a population of cells whose cycles are well synchronized. Under these admittedly optimistic circumstances, one can measure the fork density $n_d(t)$, the replication fraction $f(t)$, and averages of domain sizes. Depending on the extent of one’s *a priori* knowledge about what $I(t)$ and $v(t)$ should be and depending on the numbers and types of experiments that are possible, there are several ways to proceed. One basic issue is whether one has *a priori* knowledge about the functional form of the genome-averaged initiation rate $I(t)$ and/or that of the fork velocity $v(t)$. We outline the main possibilities below.

- (1) If the functional form is known (but not specific parameters), then one may do a least-squares curve fit to extract the unknown parameters. For example, one might suspect that $I(t) = I_n t^n$, with I_n a pre-factor and n an exponent and that v is a constant. Then one would do a curve fit to extract unknown parameters. Some programs, such as Igor Pro, support *global* curve fits where a single set of parameters (e.g., I_n , n , and v) are simultaneously fit to multiple data sets, for example to Eqs. 1 and 7. (Recall that only two among Eqs. 1, 2, 6, 7 and 8 are independent.) If global fitting is not possible, then we have found empirically that the best results to a single fit are given by fitting to the domain density, n_d (Eq. 1).
- (2) If the functional forms for $I(t)$ and $v(t)$ are unknown, then one may try to estimate these from the data. Using the results summarized in Eqs. 2–8, one can directly extract the initiation rate and fork velocity:

$$I(t) = \frac{d}{dt} \left(\frac{1}{\ell_h(t)} \right), \quad (10)$$

$$v(t) = \left(\frac{1}{2n_d(t)} \right) \frac{df}{dt}. \quad (11)$$

The latter equation can be understood as equating the growth of total domain size per length, $2vn_d$, to the rate of increase in replication fraction. One delicate point is that both these relations involve the calculation of a numerical derivative, an operation that tends to increase the effects of noise. The effects are minimized by having more data, particularly having more time points. In addition, we have found that Eq. 11 is vulnerable to systematic error at early and late replication times (e.g., before $f = 0.2$ and after $f = 0.8$). Having at least 5 time points between these two f values is essential. (Here, the issue is not only the evaluation of the numerical derivative but also that Eq. 11 assumes that the time interval

used to evaluate the derivative is short enough that no initiations or coalescences occur.)

We note, also, that the fitting and direct-inversion procedures may be combined. Starting with direct-inversion, one gets an idea of the form of either the initiation rate or fork velocity. One then guesses a functional form and uses that form as an input to the fitting procedure.

- (3) Finally, it is also possible to do independent experiments to extract the fork velocity. These would typically use a pulse-chase protocol where the nucleotide analog is added for a short time and then flushed from the experimental chamber (for example, [12]).

We illustrate the parameter-extraction procedure using *in silico* simulation data. The replication process, combing, and domain-statistics compilation are all included in the simulation. For this case, the initiation rate was assumed to increase as a power law, $I \propto t^{2.45}$, where the exponent (and prefactor) are chosen to match the values extracted from experiments on cell-free *Xenopus* embryo extracts [8]. The fork velocity was assumed to be constant (0.6 kb/min). The results are shown in Fig. 2(a)–(c), where part (a) shows the extracted averages $\bar{\ell}_i(t)$ and $\bar{\ell}_h(t)$, part (b) shows the replication fraction $f(t)$, and part (c) shows the extracted $I(t)$. Statistical errors are evaluated directly from repeated simulations; where they are not visible, they are smaller than the graph marker. At the end of S phase, errors are large because there are few domains. The solid lines are calculated from the values used to simulate the data; in particular, they are not fits. Thus, we conclude that it is possible to extract accurate estimates of replication parameters in this case.

It is worth pointing out a few details. We are essentially following the direct-inversion procedure outlined above for $I(t)$ and the fitting procedure (assuming v is constant) for the fork velocity. Figure 2(a) is thus compiled directly from experimental data. One simply measures hole and eye sizes and computes their average. (In Note 4.2, we discuss some subtleties in estimating the mean.) The error bar – present but not visible in the figure when smaller than the symbol size – is the standard error of the mean (σ/\sqrt{N} , where σ is the standard deviation of the distribution and N is the number of domains measured at a particular time point. Figure 2(b) is also calculated directly from the data. At each time point, the total length of all replicated domains is summed and then divided by the total length of all measured DNA fragments. Figure 2(c) is calculated using Eq. 10. The larger amount of statistical scatter in $I(t)$ arises from the numerical differentiation of $\bar{\ell}_h(t)^{-1}$, which tends to amplify noise.

In all the sections of Fig. 2, solid lines are calculated from the parameters used to generate the simulation. They show the good agreement between the extracted quantities and the “true” values. In this case, the solid lines are also indistinguishable from the results of least-squares fits to the data.

4 Notes

In the above discussion, our “ideal” data allowed us to successfully extract replication parameters via a simple analysis. While such data may well be obtained in the future, all experiments to date have fallen short of the criteria listed in the Materials Section. Here, we discuss how to analyze and extract parameters from data taken under the not-so-ideal conditions that, up until now, have been present. As we discuss, the significant complications have been the asynchrony of starting times for different cells and the finite length of DNA fragments that result from the combing process, and we focus on those problems. We also briefly discuss the implications of the finite (but large) length of the genome under study.

4.1 Asynchrony

Perhaps the most important limitation of experiments has been the lack of synchrony in the cell cycles of cells whose DNA was extracted for replication studies. For example, in experiments on *Xenopus* cell-free extracts, the starting time distribution had a standard deviation of 6 min, while the nominal S phase duration (10–90% replication) was 14 min. [8]. Lack of synchrony complicates the data analysis because the data from a single time point comes from a variety of actual starting times. (The asynchrony problem has one bright side: even with a small number of time points in the experiment, one probes a wide range of starting times.)

To deal with asynchrony, the basic trick is to relate all measured quantities to the replication fraction f , rather than to time [21]. In other words, one replaces the “laboratory clock” t with the “replication clock” f . Such a procedure is possible even when the cell population is completely asynchronous. Then, in a second step, whose success depends on the degree of synchrony in the starting times of S phase for the cell population under study, one converts from f to t .

The procedure begins by grouping each DNA fragment by its replication fraction f instead of by its time replicated. Having grouped fragments according to their f values, one then compiles statistics (e.g. average domain size) over each “ f -bin.” The amount of available data will determine the bin width Δf . In [8], for example, that width was a uniform 2%. In Fig. 2(d), we pooled the data from 13 groups of 100, which gave 13 f -bins of variable width. Either way is acceptable. In general, one should take bins to be as wide as possible, to minimize statistical errors, without averaging over significant features of the f dependence. It is useful to write a program that allows one to explore the effects of different choices of bin width. In any case, having settled on a

choice of bin widths, one estimates $\bar{\ell}_i(f)$, $\bar{\ell}_h(f)$, and $\bar{\ell}_{i2i}(f)$.

Once the data have been sorted by their f values, one can extract the initiation frequency I as a function of f , using expressions analogous to Eqs. 10–11, with results shown in Figs. 2(e,f):

$$\frac{I(f)}{2v} = \left(\frac{1}{\bar{\ell}_{i2i}(f)} \right) \frac{d}{df} \frac{1}{\bar{\ell}_h(f)}, \quad (12)$$

$$2vt(f) = \int_0^f \bar{\ell}_{i2i} df', \quad (13)$$

where $\bar{\ell}_{i2i}$ and $\bar{\ell}_h$ are functions of f . In other words, even for completely unsynchronized data, we can find $I(f)/2v$ vs. $2vt(f)$ from the data. At first glance, this seems to be too good to be true – up to a scale factor, one can find the form of the initiation function vs. time without any synchrony at all – but remember that what is obtained is the product $vt(f)$ (a length, which is what one measures), or $f(vt)$ if one inverts. To pass from $f(vt)$ to $f(t)$ and hence from $I(f)$ to $I(t)$ requires information that is based on the laboratory clock and not just the replication clock. This information could be obtained by making an independent measurement of the fork velocity v , as discussed previously. It is also worth pointing out that in many cases, knowing $I(f)$ is as useful as knowing $I(t)$, and thus one can gain useful information even in the absence of synchronization and without doing further experiments.

If it is important to state results in terms of the laboratory time τ and if a direct and independent measurement of v is not possible, then it is still possible to extract both v and an estimate of the distribution of replication starting times (for the different cells in the population used in experiments). One starts by constructing estimates of probability density functions (PDFs) $\rho(f, \tau_i)$ by making histograms that count the numbers of DNA fragments between f and $f + \Delta f$ that were collected at time point τ_i and then normalizing by the total number of fragments collected. In the above discussion, one would group together all fragments between f and $f + \Delta f$, regardless of time point τ_i , on the grounds that we use the f estimate of each fragment as a measure of the time at which replication started in the particular cell the fragment came from. One can then relate the set of PDFs to an unknown starting time distribution $\phi(\tau)$, which gives the proportion of cells that start replication between times τ and $\tau + \Delta\tau$:

$$\rho(f, \tau_i) = \phi(\tau) \times \left(\frac{df}{dt} \Big|_{\tau=(\tau_i-t)} \right)^{-1}. \quad (14)$$

Here, one equates $\rho(f, \tau_i)df$ with $\phi(\tau)d\tau$. In words, we observe a DNA fragment of replication fraction f at time point τ_i . Because we know the relation $f(t)$

with t the relative time elapsed since the start of replication, we can infer that this fragment came from a cell that started replicating a time t in the past, i.e., at laboratory time $\tau = \tau_i - t$. A bin of width Δf contains a fraction $\rho(f, \tau_i)\Delta f$ of the fragments that is numerically equal to $\phi(\tau)\Delta\tau$, with a width $\Delta\tau = (df/dt)^{-1}\Delta f$, where $\tau = \tau_i - t$. (Note that there are three times under discussion: t is an intrinsic clock that measures replication progress relative to the start of replication; τ is the laboratory clock; and the τ_i are particular laboratory times at which measurements are made.) We can also view Eq. 14 as a change of variables in probability distributions, from f to τ .

To return to our task of determining the velocity, we need to determine the function $\phi(\tau)$ along with the unknown velocity. It may be possible to estimate independently $\phi(\tau)$, for example by labeling newly replicated DNA [22]. If such estimates are not available, then v and $\phi(\tau)$ may be determined from the $\rho(f, \tau_i)$ by a global least-squares fit. Results are shown in Fig. 2(g,h). Alternatively, if the starting-time distribution can reasonably be approximated as Gaussian (as it could in [8]), then all that is required is an estimate of its mean and standard deviation.

Note that the uncertainty in $\phi(\tau)$ values can be estimated by transforming uncertainty estimates for $\rho(f, \tau_i)$ values. The standard way to estimate uncertainties for histogram bins is to use only bins with 5 or more instances in the bin and then to estimate the standard deviation as $\sqrt{N(f, \tau_i)}$, where $N(f, \tau_i)$ is the number of DNA fragments recorded in the bin between f and $f + \Delta f$ at time τ_i [23]. If estimates describing the shape of $\phi(\tau)$ are not available, they can be determined simultaneously with v via the above procedure. Given a candidate value for v , the derivative df/dt can be estimated, for example, from the previously determined $vt(f)$. Of course, determining more parameters from the data will increase the uncertainty of the resulting estimates.

One final subtlety is that fragments with $f = 0$ or 1 are problematic. For example, we cannot infer a unique starting time to an $f = 1$ fragment, since a cell will stay at $f = 1$ for a long period after finishing replication. We thus exclude these fragments from the analysis.

4.2 Finite Fragment Sizes

The second potential complication arises from the finite size of DNA fragments. The first generation of replication combing experiments (e.g. [2]) gave DNA fragments averaging about 200 kb in length. The average fragment length is determined by factors such as the fluid-flow shear associated with molecular combing, with the largest fragments exceeding 1 Mb [24]. As mentioned in the Materials Section, the absolute length of fragment is not important; rather,

what counts is the number of domains per fragment. From Eq. 1, one can show that this number is low at the beginning and end of S phase and reaches a maximum in the middle of S phase. Thus, while a minimal requirement for a successful experiment is that there exist a reasonable range of f values where the typical DNA fragment has many (say 10) domains, any experiment will have problems at the beginning ($f \rightarrow 0$), where the average hole size on the original, unbroken chromosome will eventually exceed the average fragment size and the end ($f \rightarrow 1$), where the average eye size will eventually exceed the average fragment size.

The simplest way to deal with this problem is to simply ignore all DNA fragments that have fewer than some minimal number (say 5) of domains. While such a rule of thumb keeps the uncertainty of estimated parameters bounded, it implies that little information will be gathered about the first and last stages of replication. In order to increase the information extracted from experiments in those regimes, one can do a more sophisticated analysis [25]. This analysis begins by recognizing that there are three classes of domains (either holes or eyes): interior, exterior, and over-sized (Fig. 3). Up to now, we have implicitly assumed that all domains were interior domains. An interior eye, for example, is one that is flanked by two hole domains, allowing its size to be measured unambiguously. An edge-eye domain is bounded on one side by a hole domain and on the other by the edge of the molecule. Thus, one cannot know the true size of the eye domain as it existed on the original, unbroken chromosome. The worst case is that of an oversized domain, where the domain extends beyond both edges of the DNA fragment, Fig. 3(b). One can picture the situation as one where an initial distribution of, say, eye sizes is subdivided into three experimental distributions of interior, edge, and oversized domain lengths. The problem, then, is that the naive estimator of average eye size,

$$\bar{\ell}_{interior} = \frac{L_{interior}}{N_{interior}}, \quad (15)$$

(the total length of interior domains divided by their total number) is biased. Intuitively, it must always be smaller than the true value because some large domains will show up as edge or oversized domains. Because of the direct role of average domain sizes in our analysis, any bias in those quantities will bias the inferred initiation and fork rates.

If the population is well-synchronized, one can show that it is possible to construct an *unbiased* estimator of the average domain size,

$$\bar{\ell}_{unbiased} = \frac{L_{total}}{N_{total}} = \frac{L_{interior} + L_{edge} + L_{oversized}}{N_{interior} + N_{edge}/2}, \quad (16)$$

where $L_{total} = L_{interior} + L_{edge} + L_{oversized}$ is the total length of all fragments

analyzed and $N_{total} = N_{interior} + N_{edge}/2$ is the total number of domains in the unfragmented DNA, equal to the number of interior and half the edge fragments. (The factor of 1/2 arises because each time the original DNA molecule breaks, two edge domains are produced. Note that oversized domains do not contribute). In practice, an experiment will likely show effects from finite fragment sizes *and* asynchrony. This poses a problem for the previous analysis, as it is no longer possible to determine which f value to assign a given oversized fragment. Still, one can show that the obvious work-around — simply to omit $L_{oversized}$ from Eq. 16 — reduces the bias of the naive estimator $\ell_{interior}$ for domain size by including information about the edge domains [25]. For more accurate results, then, one should use this “interior-edge” estimator.

4.3 Finite Genome Length

So far, we have implicitly assumed that the total length of the genome is infinite. This is apparent in the expression for $f(t)$, where, for example, Eq. 3 implies that $f \rightarrow 1$ as $t \rightarrow \infty$, meaning that it takes an infinite amount of time to complete S phase ($f = 1$). But obviously, a finite genome replicates in a finite time.

For most practical measurements, the genome is so large that the differences between the infinite-genome approximation and the finite-genome result are very small. For example, if one calculates the time to go from 5 to 95% replicated (i.e., from $f = 0.05$ to 0.95), the infinite-genome result will not be measurably different. However, in certain cases, it is important to be able to calculate the exact duration of S phase (i.e., from $f = 0$ to 1). For example, in *Xenopus* embryos before the mid-blastula transition, the duration of S phase is about 20 min. while the entire cell cycle is only 25 min. [26]. For such a case, it can be interesting to be able to infer the duration of S phase indirectly from measurements throughout the synthesis phase. Here, we summarize results from a recent theoretical study of this case [10].

In a finite genome, the stochasticity (randomness) of initiation will imply that the duration of S phase is also a stochastic variable and will thus have an “end-time” distribution $\rho_{end}(t)$. The mean of this distribution gives the average time to replicate the entire genome. Its standard deviation gives the typical variation in this time, which can be taken as a measure of the “reliability” of the replication process and the need for checkpoint mechanisms to compensate. For example, in the example of *Xenopus* embryos given above, the reliability must be high — and σ correspondingly small — in order for replication to be complete before the end of the cell cycle. If replication is not complete by the end of mitosis, “mitotic catastrophe” ensues [10].

Using methods of extreme-value statistics, one can show that $\rho_{end}(t)$ is approximately a Gumbel distribution, $\exp(-x) \exp(-e^{-x})$, where $x = (t - t^*)/\beta$ is a dimensionless variable [27]. The location parameter t^* gives the mode of the distribution, and the scale parameter β gives the width. The mean $t_{avg} = t^* + \beta\gamma$, with $\gamma = 0.57721\dots$ the Euler-Mascheroni constant. The standard deviation $\sigma = (\pi/\sqrt{3})\beta \approx 1.2825\beta$.

The next step is to determine t^* and β in terms of the fork velocity v , initiation rate $I(t)$, and chromosome length L . The mode t^* is determined via an implicit transcendental equation [10],

$$Lg(t^*) \exp \left[-2v \int_0^{t^*} g(t') dt' \right] = 1, \quad (17)$$

where $g(t) = \int_0^t I(t') dt'$ has units of (1/length) since $I(t)$ is the number of initiations per time per length. One can solve Eq. 17 numerically using a standard one-dimensional nonlinear equation solver, such as *FindRoots* in IgorPro, to find t^* . The width of the end-time distribution is given by $\beta = 2vg(t^*)$.

4.4 Combined analysis

Finally, we present the results of an analysis of simulated data that includes all of the issues discussed above (Fig. 2)(d)–(h). The simulations are done using the same parameter values as used in (a)–(c). The difference is that now there is a population of 100 cells, whose replication starting time is drawn from a Gaussian distribution. We sample 13 times from each cell, with each fragment 1 Mb long. In (d), we present the average domain size as a function of f . These are directly measured from the data. In (e), we present f as a function of $2vt$, with v the (so-far unknown) fork velocity. The calculation is done using Eq. 13. Similarly, we estimate the quantity $I/2v$ vs. $2vt$ using Eq. 12. At this point, v is still unknown, but we can step through a set of possible values. For each, we sum the squares of the deviations between the measured and predicted values (via Eq. 14) of each bin of the $\rho(f, t_i)$ histograms, which gives us the χ^2 statistic [28]. Because we calculate a single χ^2 statistic by summing over all the $\rho(f, t_i)$ histograms, this is a global fit. The minimum value of $\chi^2(v)$, 0.596 ± 0.039 kb/min, is consistent with the 0.6 kb/min used in the simulations. Given a velocity, one can then work out the starting-time distribution $\phi(t)$, shown here in Fig. 2(h). From that, one can calculate $f(t)$ and $I(t)$. The new axes (just rescalings of $2vt$) are shown as top and right axes in Fig. 2(e,f). We conclude that reasonable inferences about the fork velocity, initiation rate, and related quantities can be made even in the presence of “real-life” experimental issues.

Acknowledgements

I thank my former students Suckjoon Jun, Haiyang Zhang, and Brandon Marshall for all their contributions to the development of the methods described here. I thank Aaron Bensimon and John Herrick for their collaboration and for having introduced me to this fascinating area of science. I thank Nick Rhind and John Herrick for their comments on a draft of this chapter. This work was supported by an NSERC Discovery Grant (Canada) and by the Human Frontier Science Program.

References

- [1] Bensimon A, Simon A, Chiffaudel A, Croquette V, Heslot F, Bensimon D. Alignment and sensitive detection of DNA by moving interface. *Science* 1994;264:2096–8.
- [2] Herrick J, Stanislawski P, Hyrien O, Bensimon A. Replication fork density increases during DNA synthesis in *X. laevis* egg extracts. *J Mol Biol* 2000;300:1133–42.
- [3] Norio P, Schildkraut CL. Visualization of DNA replication on individual Epstein-Barr virus episomes. *Science* 2001;294:2361–4.
- [4] Pasero P, Bensimon A, Schwob E. Single-molecule analysis reveals clustering and epigenetic regulation of replication origins at the yeast rDNA locus. *Genes & Dev* 2002;16:2479–84.
- [5] Anglana M, Apiou F, Bensimon A, Debatisse M. Dynamics of DNA replication in mammalian somatic cells: Nucleotide pool modulates origin choice and interorigin spacing. *Cell* 2003;114:385–94.
- [6] Patel PK, Arcangioli B, Baker SP, Bensimon A, Rhind N. DNA replication origins fire stochastically in fission yeast. *Mol Biol Cell* 2006;17:308–16.
- [7] Di Micco R, Fumagalli M, Cicalese A, Piccinin S, Gasparini P, Luise C, Schurra C, Garré M, Nuciforo PG, Bensimon A, Maestro R, Pelicci PG, d’Adda di Fagagna F. Oncogene-induced senescence is a DNA damage response triggered by DNA hyper-replication. *Nature* 2006;444:638–42.
- [8] Herrick J, Jun S, Bechhoefer J, Bensimon A. Kinetic model of DNA replication in eukaryotic organisms, *J Mol Biol* 2002;320:741–50.
- [9] Hyrien O, Marheineke K, Goldar A. Paradoxes of eukaryotic DNA replication: MCM proteins and the random completion problem. *BioEssays* 2003;25:116–25.
- [10] Bechhoefer J, Marshall B. How *Xenopus laevis* replicates DNA reliably even though its origins of replication are located and initiated stochastically. *Phys Rev Lett* 2007;98:098105:1–4.

- [11] Jun S, Herrick J, Bensimon A, Bechhoefer J. Persistence length of chromatin and origin spacing in *Xenopus* early embryo DNA replication. *Cell Cycle* 2004;3:223–9.
- [12] Conti C, Saccà B, Herrick J, Lalou C, Pommier Y, Bensimon A. Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. *Mol Biol Cell* 2007;18:3059–67.
- [13] Marheineke K, Hyrien O. Control of replication origin density and firing time in *Xenopus* egg extracts: Role of a caffeine-sensitive, ATR-dependent checkpoint. *J Biol Chem* 2004;279:28071–81.
- [14] Shechter D, Gautier J. ATM and ATR check in on origins. *Cell Cycle* 2005;4:235–8.
- [15] Herrick J, Bensimon A. Global regulation of genome duplication in eukaryotes: an overview from the epifluorescence microscope. *Chromosoma* 2008; in press.
- [16] Kolmogorov AN On the statistical theory of crystallization in metals. *Izv Akad Nauk SSSR, Ser Fiz* 1937;1:355–9.
- [17] Johnson WA, Mehl RF. Reaction kinetics in processes of nucleation and growth. *Trans AIME* 1939;135:416–42. Discussion 442–58.
- [18] Avrami M. Kinetics of Phase Change. I. General theory. *J Chem Phys* 1939;7:1103–12. Kinetics of Phase Change. II. Transformation-time relations for random distribution of nuclei. *Ibid* 1940;8:212–24. Kinetics of phase change III. Granulation, phase change, and microstructure. *Ibid* 1941;9:177–84.
- [19] Diffley JF. Once and only once upon a time: specifying and regulating origins of DNA replication in eukaryotic cells. *Genes & Dev* 1996;10:2819–30.
- [20] Jun S, Zhang H, Bechhoefer J. Nucleation and growth in one dimension, part I: The generalized Kolmogorov-Johnson-Mehl-Avrami model. *Phys Rev E* 2005;71:011908:1–8.
- [21] Blumenthal AB, Kriegstein HJ, Hogness DS. The units of DNA replication in *Drosophila melanogaster* chromosomes. In: *Cold Spring Harbor Symp on Quant Biol* 1974;38:205–23.
- [22] Schermelleh L, Solovei I, Zink D, Cremer T. Two-color Fluorescence labeling of early and mid-to-late replicating chromatin in living cells. *Chrom Res* 2001;9:77–80.
- [23] Sivia DS. *Data Analysis: A Bayesian Tutorial*. 2nd ed. Oxford, England: Oxford University Press, 2006.
- [24] Michalet X, Ekong R, Fougerousse F, Rousseaux S, Shurra C, Hornigold N, van Slegtenhorst M, Wolfe J, Povey S, Beckmann JS, Bensimon A. Dynamic molecular combing: stretching the whole human genome for high-resolution studies. *Science* 1997;277:1518–23.

- [25] Zhang H, Bechhoefer J. Reconstructing DNA replication kinetics from small fragments. *Phys Rev E* 2006;73:051903:1–9.
- [26] Hyrien O, Méchali M. Chromosomal replication initiates and terminates at random sequences but at regular intervals in the ribosomal DNA of *Xenopus* early embryos. *EMBO J* 1993;12:4511–20.
- [27] Gumbel EJ. *Statistics of Extremes*. New York, NY: Columbia University Press, 1958.
- [28] Jun S, Bechhoefer J. Nucleation and growth in one dimension, part II: Application to DNA replication kinetics. *Phys Rev E* 2005;71:011909:1–8.

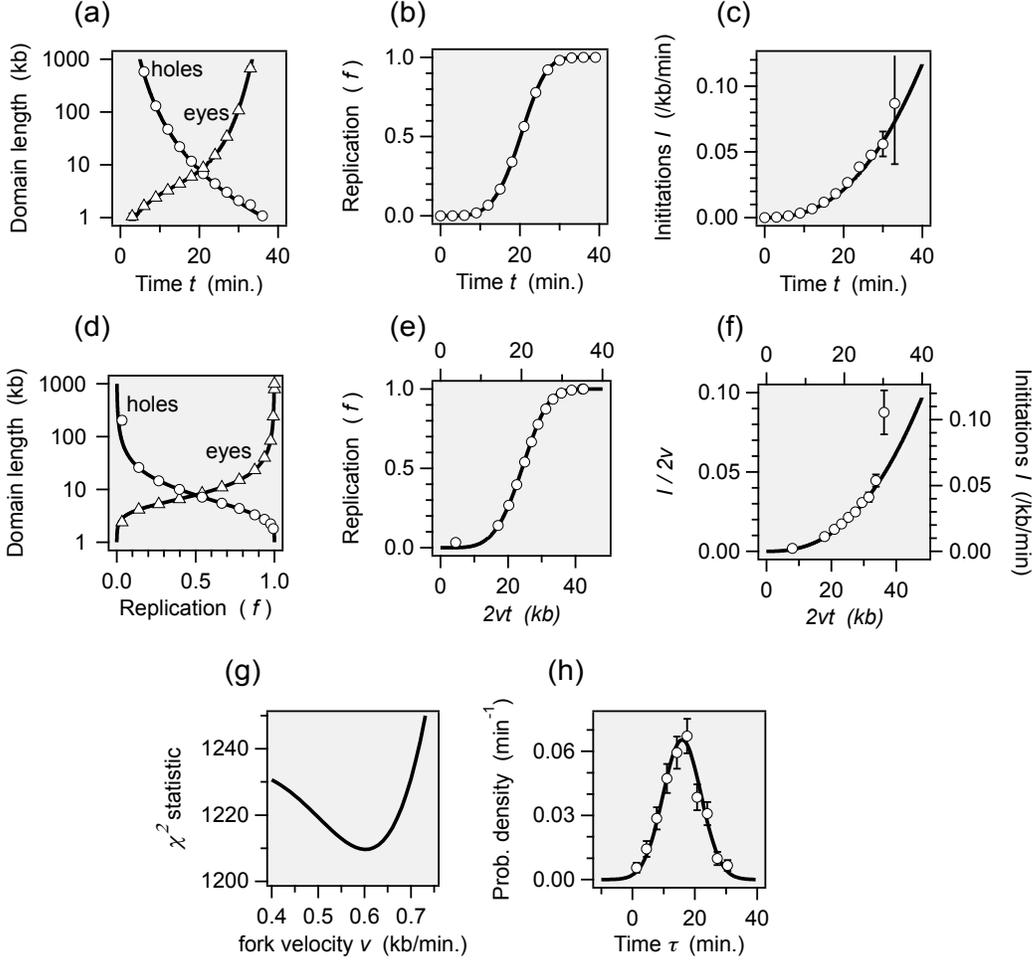


Fig. 2. Parameter extraction from almost ideal and more realistic simulated data sets. In all cases, the thick solid lines correspond to the parameters actually used in simulating the data – they are not fits. The parameters ($I(t) = I_n t^n$ /min/kb, with $I_n = 1.38e-5$, $n = 2.45$, and $v = 0.6$ kb/min) were chosen to correspond to those found for *Xenopus* cell-free embryo extracts [8]. Errors are estimated by compiling statistics from repeated simulations. (a)–(c) Analysis of an almost ideal data set of length 100 Mb, chopped into fragments 1 Mb long, with 13 time points taken at intervals of 3 min. Data are perfectly synchronous. (a) Average eye and hole domain sizes vs. time. (b) Replicated fraction vs. time. (c) Inferred initiation rate vs. time. (d)–(h) Analysis of a more realistic data set also consisting of 13 time points where 100 samples, each 1 Mb long, are taken from a population of 100 cells. The starting times of replication of the 100 cells are drawn from a Gaussian distribution with a standard deviation of 6.1 min. Otherwise, the same parameters are used as above. (d) Average eye and hole domain sizes vs. replication fraction f . (e) Replication fraction f vs. $2vt$ (bottom axis). After v is determined, the $2vt$ axis may be rescaled in terms of t alone (top axis). (f) Scaled origin initiation rate $I/2v$ vs. $2vt$. Again, after determining v , one can rescale axes in terms of I vs. t (right and top axes). (g) The minimum value of the χ^2 statistic gives the fork velocity. (h) Starting-time distribution $\phi(\tau)$.

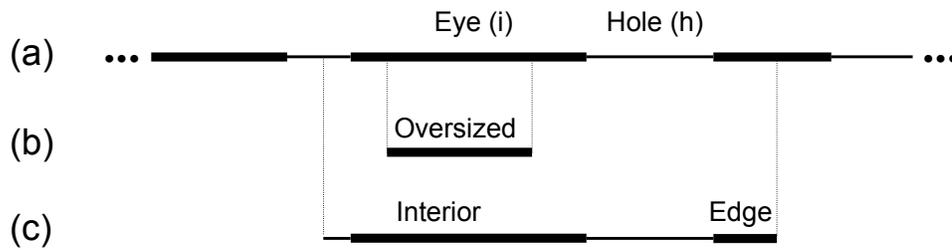


Fig. 3. Sketch of the three types of eye domains. (a) Portion of a very long DNA fragment showing eye and hole domains. (b) Short fragment consisting of an oversized eye domain. (c) Longer fragment with interior and edge eye domains indicated.