# 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*

antibody attached to a di erent-color uorophore, 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 di erent 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 nal quantity of interest is the \eye-to-eye" distance, de ned 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.openo ce.org). Alternatively, a more-sophisticated scienti c 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 tting to experimental data. This can be important in that deriving t17ki6(ey)2o1 hatnItalticat(nl)-8m-27(ducs)-44Td [m28(v)y44Td [-27(e)-393wneet17tTJ 0 -14

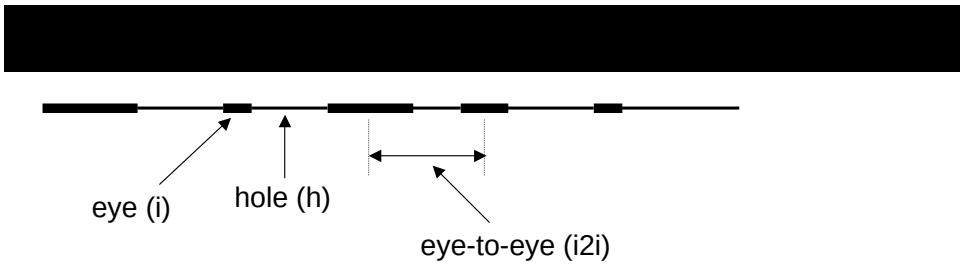


Fig. 1. Top: Epi uorescence image of a combed fragment of DNA labeled to show non-replicated areas. Non-replicated segments are visualized using anti-BrdU antibodies. The length and continuity of the DNA fragment is determined by labeling with anti-guanosine antibodies (image not shown). Bottom: Schematic 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 nite 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 nite-size e ects are small.

(3) Good optical resolution and good labeling e ciency 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-speci c 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

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 cuto have just a single origin.

The kinetic-modeling approach presented here skirts these di culties. Because the model is statistical, it can incorporate all the acquired data. In e ect, 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

| symbol | de nition |
|---|---|
| $f$ | 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 |
| $`_i$ | average length of replicated domains (\eyes") |
| $`_h$ | average length of non-replicated domains (\holes") |
| $`_{i2i}$ | 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 |
| $`_{interior}$ | biased domain-length estimator using only interior domains |
| $`_{unbiased}$ | unbiased domain-length estimator from interior, edge, and oversized domains |
| $t$ | time elapsed since start of replication |
| | laboratory time |
| $_i$ | times at which replication data are collected |
| $( )$ | distribution of starting times of DNA replication for di erent cells |
| $(f; _i)$ | distribution of replication-fraction values of DNA fragments collected at time $_i$ |
| $_{end}(t)$ | distribution of replication times for a nite genome |
| $t^*$ | typical time to replicate completely a genome (mode of end-time distribution) |
| | width (in time) of end-time distribution ( / 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 $I(t) = I_0$,

### 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 e cient and speci c 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 speci c parameters), then one may do a least-squares curve t 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 t to extract unknown parameters. Some programs, such as Igor Pro, support *global* curve ts where a single set of parameters (e.g., $I_n$, $n$, and $v$) are simultaneously t 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 tting is not possible, then we have found empirically that the best results to a single t are given by tting 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:

$$I(t) = \frac{d}{dt}\left[\frac{1}{\lambda_h(t)}\right]; \tag{10}$$

$$v(t) = \frac{1}{2n_d(t)}\frac{df}{dt}: \tag{11}$$

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 e ects of noise. The e ects 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 tting 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 tting 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 ushed 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 $\grave{}_i(t)$ and $\grave{}_h(t)$, 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 ts. Thus, we conclude that it is possible to extract

## 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 signi cant complications have been the asynchrony of starting times for di erent cells and the nite length of DNA fragments that result from the combing process, and we focus on those problems. We also brie y discuss the implications of the nite (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

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

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):

$$\frac{I(f)}{2v} = \frac{1}{\grave{}_{i2i}(f)} \left( \frac{d}{df} \frac{1}{\grave{}_h(f)} \right) ; \tag{12}$$

$$2vt(f) = \int_0^f \grave{}_{i2i} df' ; \tag{13}$$

where $\grave{}_{i2i}$ and $\grave{}_h$ are functions of $f$. In other words, even for completely unsynchronized data, we can nd $I(f)=2v$ vs. $2vt(f)$ from the data. At rst glance, this seems to be too good to be true { up to a scale factor, one can nd 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 $f$

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 $= {}_i \quad t$. A bin of width $f$ contains a fraction $(f; {}_i) \quad f$ of the fragments that is numerically equal to $( ) \quad$, with a width $= (df=dt)^{-1} \quad f$, where $= {}_i \quad t$. (Note that there are three times under discussion: $t$ is an intrinsic clock that measures replication progress relative to the start of replication; is the laboratory clock; and the ${}_i$ 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 .

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 ($f \; ! \; 0$), where the average hole size on the original, unbroken chromosome will eventually exceed the average fragment size and the end ($f \; ! \; 1$), 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,

$$\grave{}_{interior} = \frac{L_{interior}}{N_{interior}} \; ; \tag{15}$$

(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,

$$\grave{}_{unbiased} = \frac{L_{total}}{N_{total}} = \frac{L_{interior} + L_{edge} + L_{oversized}}{N_{interior} + N_{edge}/2} \; ; \tag{16}$$

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 e ects from nite fragment sizes *and* asynchrony. This poses a problem for the previous analysis, as it is no longer possible to determine which $f$

## 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, Chi audel 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 re stochastically in ssion yeast. Mol Biol Cell 2006;17:308{16.

[7] Di Micco R, Fumagalli M, Cicalese A, Piccinin S, Gasparini P, Luise C, Schurra C, Garre 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]

[25] Zhang H, Bechhoefer J. Reconstructing DNA replication kinetics from small fragments. Phys Rev E 2006;73:051903:1{9.

[26] Hyrien O, Mechali 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.

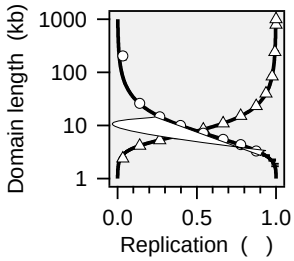


Fig. 2. Parameter extraction from almost ideal and more realistic simulated data sets. In all cases, the thick solid lines correspond to the parameters actually used in simulating the data { they are not ts. The parameters ($I(t) = I_n t^n$ /min/kb, with $I_n$ = 1:38e-5, $n$ = 2:45, and $v$ = 0:6 kb/min) 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 $^2$ statistic gives the fork velocity. (h) Starting-time distribution ( ).

Eye (i)

(a)