MODELLING THE DNA REPLICATION PROGRAM IN EUKARYOTES

by

Scott Cheng-Hsin Yang

B.Sc. (Hons., Physics), University of British Columbia, 2006

THESIS SUBMITTED IN PARTIAL FULFILLMENT

OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN THE

DEPARTMENT OF PHYSICS

FACULTY OF SCIENCE

© Scott Cheng-Hsin Yang 2012

SIMON FRASER UNIVERSITY
Spring, 2012

All rights reserved. However, in accordance with the Copyright Act of Canada, this work may be reproduced, without authorization, under the conditions for Fair Dealing. Therefore, limited reproduction of this work for the purposes of private study, research, criticism, review, and news reporting is likely to be in accordance with the law, particularly if cited appropriately.

APPROVAL

Name: Scott Cheng-Hsin Yang

Degree: Doctor of Philosophy

Title of Thesis: Modelling the DNA Replication Program in Eukaryotes

Examining Committee: Dr. J. Steven Dodge (Chair)

Dr. John Bechhoefer
Senior Supervisor
Professor

Dr. Martin Zuckermann
Supervisor
Adjunct Professor

Dr. Nicholas Rhind
Supervisor
Associate Professor of Biochemistry
& Molecular Pharmacology
University of Massachusetts
Medical School, Worcester, MA

Dr. Levon Pogosian
Internal Examiner
Assistant Professor

Dr. Eldon Emberly
Supervisor
Associate Professor

Dr. John F. Marko
External Examiner
Professor of Physics & Astronomy and
Professor of Molecular Biosciences
Northwestern University, Evanston, IL

Date Approved: April 4, 2012

Partial Copyright Licence

The author, whose copyright is declared on the title page of this work, has granted to Simon Fraser University the right to lend this thesis, project or extended essay to users of the Simon Fraser University Library, and to make partial or single copies only for such users or in response to a request from the library of any other university, or other educational institution, on its own behalf or for one of its users.

The author has further granted permission to Simon Fraser University to keep or make a digital copy for use in its circulating collection (currently available to the public at the "Institutional Repository" link of the SFU Library website (www.lib.sfu.ca) at http://summit/sfu.ca and, without changing the content, to translate the thesis/project or extended essays, if technically possible, to any medium or format for the purpose of preservation of the digital work.

The author has further agreed that permission for multiple copying of this work for scholarly purposes may be granted by either the author or the Dean of Graduate Studies.

It is understood that copying or publication of this work for financial gain shall not be allowed without the author's written permission.

Permission for public performance, or limited permission for private scholarly use, of any multimedia materials forming part of this work, may have been granted by the author. This information may be found on the separately catalogued multimedia material and in the signed Partial Copyright Licence.

While licensing SFU to permit the above uses, the author retains copyright in the thesis, project or extended essays, including the right to change the work for subsequent purposes, including editing and publishing the work in whole or in part, and licensing other parties, as the author may desire.

The original Partial Copyright Licence attesting to these terms, and signed by this author, may be found in the original bound copy of this work, retained in the Simon Fraser University Archive.

Simon Fraser University Library
Burnaby, British Columbia, Canada

revised Fall 2011

Abstract

DNA replication in higher organisms starts at many places across the genome and throughout S (synthesis) phase. In order to understand replication in eukaryotes, one needs to know not only how the replicative machinery functions on the molecular level but also how the machinery is organized genome wide to ensure complete duplication. Over the past fifteen years, advances in technology have allowed researchers to perform genome-wide experiments that probe the state of replication in many organisms. These datasets make possible quantitative modelling of the replication process.

The kinetics of DNA replication is formally analogous to a physical phase-transformation process. In replication, the DNA is transformed from a "non-replicated" phase to a "replicated" phase, just as freezing water is transformed from a liquid phase to a solid phase. Using this analogy, we map the replication process onto a stochastic nucleation-and-growth model introduced in statistical physics to describe first-order phase transitions. Extending the model, we develop a mathematical framework that is flexible enough to describe the kinetics of replication in eukaryotes.

We present three applications of our theory: 1) We apply the theory to a recent dataset on budding yeast to extract its genome-wide replication program. Based on this study, we give the first proposal to explain how the temporal aspect of the replication program can be controlled mechanistically. 2) We address the "random-completion problem," which asks how replication-completion times can be controlled when replication starts at random places and times. We find that the strategy adopted in frog embryos to solve the problem also nearly minimizes the use of certain replicative machinery. 3) We study possible ways to extract information from a popular technique used to probe replication in multicellular eukaryotes, ranging from worms to humans. We show preliminary results that can be extended to real experiments in the near future.

給 媽

Acknowledgments

Throughout the six years of my PhD, I imagined many different starting points and turn outs. Each time, I came to the conclusion that I am most fortunate to have John as my supervisor. There are too many reasons. To start, John is fast with emails: he'll reply my e-mail in less than a few hours if I write him one now. This habit of his shows that he is always available and ready to help. John is relentlessly patient, especially in perfecting my writing. For our first publication, he must have gone through my paper close to a hundred times, helping me with my never-ending spelling, grammar, and punctuation mistakes one by one. John is understanding. I believe he understands my work habits and guides me with the right amount of freedom and pressure. He also understands that there are more important things than research in my life and is considerate in those respects. John is foresightful. Many times, I would see a book from him on my lab desk. The books covered a diverse range of topics, from punctuation to making graphs, from data analysis to mastering PhD. He would also push me to attend conferences and give presentations. After all these years, I am starting to understand why John did all these. John is knowledgeable. To quote my fellow graduate students: "John is scary. He seems to know everything." Indeed, I keep on being surprised by the breadth and depth of John's knowledge. Just recently, while playing with a demonstration on artificial rainbow, John commented on how the optics of rainbow formation connects to a theory in particle physics. Amazing eh? Most notably, John is fun to work with. We had countless discussions where I ended up laughing really hard. It is my true pleasure and honour to work with John. Thanks a lot, John.

I also must acknowledge my SFU connections: thanks to Nick, who supported me in everything I asked for, to Michel, who became to me like an elder brother, to Yiwei and Jixin, my true friends (very difficult to find!), to Suckjoon, who was legendary in my dinner-table conversations, to Levon, in whose course I finally started to enjoy learning physics

without fear, to Paul, who opened the field of machine learning to me, to Eldon, Martin, Malcolm, Steve, Nancy, ChangMin, Yonggun, our department staff Jen, Amy, Rose, and many others. These people made my years at SFU so wonderful that I have the courage to take my next step toward an academic profession.

Of course, I will be nowhere without my parents' unconditional love and care. I thank my Dad for always providing me with inspiring ideas and supporting me in every choice I make. Special thanks to my dearest Mom for taking good care of me, for respecting my imagination, for teaching me that life (and therefore science) is supposed to be simple, and for leaving me a wealth of memories from which I have been and will keep enjoying and learning. This thesis is dedicated to her; I know she is proud of me, as always.

I thank Sherry my bride-to-be, who is beautiful in and out. She is always there for me, at good and bad times. My PhD years would be less than half as rewarding without her company.

Special thanks to my homestay family, the Lin family (Uncle Peter, Auntie Grace, Ellen, and Doris). Although it took me 1.5 hours to commute one way from home to SFU, I did not move out after moving in around 2004. The nine years I have spent in this home are unforgettable.

I owe the most to my Lord Jesus Christ, who is my all-sufficient grace, my Emmanuel Shepherd, the True, the incomparably precious One, and the cherishing and trimming Son of Man. I love His purpose—to gain a group of people who are one with Him in life, nature, person, function, and expression, but not in the Godhead. Although the content of this thesis is not directly relevant to His purpose, the discipline and experiences needed to complete the thesis are in accordance with His purpose. I honour His ways and praise Him for Who He is and for that He is.

Life would be black and white without my brothers and sisters in the church. Thanks to Wayne, whom I use as a proof to show people that I have good friends. Thanks to Shane, who prayed with me on his knees. Thanks to Frank, Louie, Andy, Robin, Eric, Nick, Selow, and many more brothers for living the church life with me. Thanks to all the uncles and aunties who served me: they cherished, nourished, and showed me what are the most important and worthwhile things in life. Thanks to Doris and Ellen, whose day-to-day care is heart warming. Thanks to Nini and Jessica, who are my patterns in their genuine and selfless care for others. Thanks to all the brothers and sisters who cared and prayed for me.

They might not be directly involved in my studies, but it is because of them that I can come this far.

Lastly, thanks to my dear last-minute proofreaders (I was last minute; they were not) Kabs, Yuli, Hilary, Doris, Ellen, James, Wayne, Shane, Frank, and Ricky.

Unfortunately, I cannot list every name on my heart at this moment. But I am sincerely thankful that I am part of their lives and they are part of mine.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Faithful and timely DNA replication is essential for the normal development of life. Unfaithful and uncontrolled replication of the genome—for example, mis-replication, re-replication, and partial replication—can lead to chromosomal instability that activates programmed cell death or oncogenes [1, 2]. Decades of intense research has revealed the key molecular players and biochemical processes [3, 4, 5]. This understanding has led to a complete description of the replication kinetics in bacteria: DNA synthesis starts at a single, sequence-specific site, proceeds bidirectionally from it, and finishes the duplication at another sequence-specific region [6]. In this case, the genome-wide regulation of the replication process is deterministic and strictly governed by biochemical reactions. (The reactions themselves, of course, are stochastic in this and all cases.)

Eukaryotic replication is more complex. The kinetics depends not only on the properties of the replicative proteins but also on their genome-wide coordination and organization [7]. Over the past fifteen years, advances in technology have allowed genome-scale experiments that probe the state of replication. With such datasets comes the potential to construct a complete and detailed picture of how replication occurs and is regulated [8]. In this thesis, I will develop a mathematical formalism that describes the eukaryotic replication kinetics, facilitates logical information extraction from experiments, and addresses apparent contradictions and robustness issues arising from the stochastic nature of replication.

1.1 Overview of replication

Common to all organisms is their unique ability to reproduce. This is true for the most basic building block of an organism, a cell. The life of a cell, from its "birth" to its duplication, is governed by a cell cycle (Fig. 1.1). The cell cycle for somatic cells is composed of four phases: the Gap 1 (G1) phase, where cells grow in size and prepare the necessary biochemical environment for DNA replication; the synthesis (S) phase, where DNA synthesis takes place to duplicate the genome; the Gap 2 (G2) phase, where cells continue to grow and prepare the environment for dividing; and the mitosis (M) phase, where a cell divides into two cells, each carrying a copy of the original genome. The embryonic cell cycle is shortened with the omission of G1 and G2 phases. By stockpiling the necessary proteins in a large egg, the organism can develop more quickly.

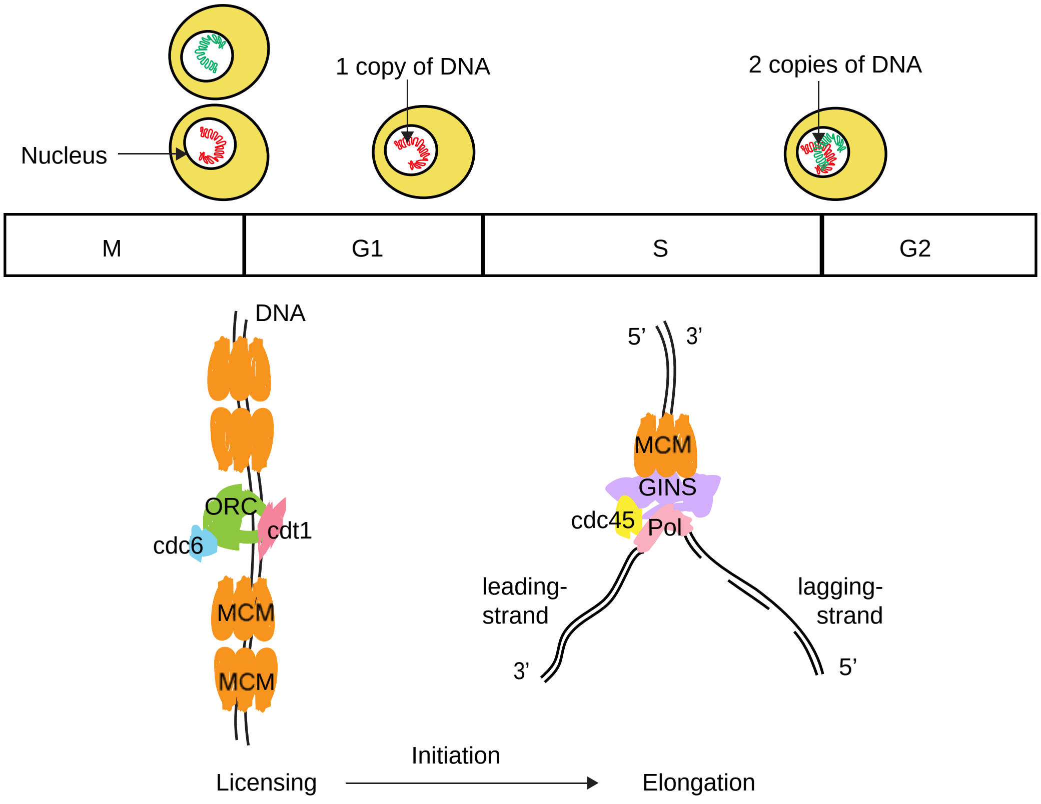


Figure 1.1: A schematic diagram of the cell cycle and the major molecular processes in DNA replication.

DNA is composed of four kinds of bases, or nucleotides: A (adenine), T (thymine), C (cytosine), and G (guanine). The DNA that encodes genetic information in a cell's nucleus is in a double-stranded form, where an A/C on one strand is complemented by a T/G on the other strand. The matching of the letters is called complementary base pairing. In order to duplicate a molecule of double-stranded DNA (dsDNA), the dsDNA is unwound into two single-stranded DNA (ssDNA). Replication proteins then bind the ssDNA to synthesize a new complementary second strand until the whole molecule is copied. Below, we explain in more detail the major steps involved in DNA replication [4, 5]:

> **Licensing**. Licensing establishes "potential origins" along the DNA, which can initiate replication. The licensing process involves the formation of pre-replicative complexes (pre-RC). Each complex is first formed when a single group of six proteins, the origin recognition complex (ORC), binds to the DNA. Each ORC, with the help of two additional proteins (Cdc6 and Cdt1) recruits minichromosome maintenance (MCM) 2-7 hexamer rings onto the chromosome (Fig. 1.1). A pair of head-to-head MCM rings can function as a potential origin [9, 10]. Licensing in somatic cells is restricted to M and G1 phase of the cell cycle by the requirement of low cyclin-dependent kinase (CDK) activity.
>
> **Initiation**. After licensing and upon entering S phase, a potential origin can be activated by the phosphorylation of CDK and Dbf4-dependent kinase (DDK) and by the recruitment of various proteins, such as Cdc45 and the GINS complex. Once an origin is activated, the pre-RC disassembles. The pair of MCM2-7 rings, with Cdc45 and GINS, moves bidirectionally outward from the origin as helicases to unwind the double-stranded DNA, forming two symmetrically propagating replication "forks" (Fig. 1.1). The term "fork" is used at the unwinding front because the geometry of the dsDNA unzipped into two ssDNA is fork-shaped. Throughout the thesis, the verbs "activate," "initiate," and "fire" are used to describe this event. Usually, these three terms are equivalent; however, in Chapter 4, a distinction needs to be made. The word "activate" is associated with an activator, which is freely diffusing in the nucleus, while the words "initiate" and "fire" are associated with an origin, which is on the DNA.

Elongation[*]. Following the unwinding of the dsDNA into ssDNA, polymerases are recruited. Polymerases can synthesize nucleotides in only one direction, from the $5'$ end of the DNA to the $3'$ end. Since the newly synthesized strand is complementary to the old, the old strand that is oriented in the $3'$–$5'$ direction (leading-strand) is replicated continuously. The other strand, which is oriented in the $5'$–$3'$ direction (lagging-strand), is replicated discontinuously with Okizaki fragments (Fig. 1.1). Reviews of the important proteins involved can be found in [11, 12].

Coalescence. When two replication forks travelling in opposite directions meet, the helicases disassemble, and the two growing strands of newly synthesized DNA are joined together by DNA ligases.

In eukaryotic cells, the processes of origin initiation, fork progression (with tightly coupled replication machinery), and domain coalescence take place at multiple sites throughout S phase until the whole genome is duplicated. Re-replication is prevented because pre-RCs are licensed only in M and G1 phases and not in S phase. (If licensing were allowed in S phase, the pre-RCs could be licensed on replicated DNA to re-replicate it.) When potential origins initiate or are passively replicated by other replication forks[†], pre-RCs disassemble and are inhibited from reassembling on the DNA throughout the current S phase, thereby preventing re-initiation and re-replication [4].

As mentioned previously, the replication process depends not only on the proper functioning of the protein complexes but also on how they are organized. More specifically, the positional organization of potential origins in licensing, the temporal organization of initiation in S phase, and the velocity of the fork movements need to be regulated to ensure complete and timely duplication of the genome. The collection of these elements and the resulting coalescences are referred to as "replication kinetics." As we will discuss later, the work presented here focuses on a scenario where the fork velocity is constant. In this case, the replication kinetics is essentially determined by the licensing position and the initiation time. We refer to this spatiotemporal organization as the "replication program."

[*] An animation of the elongation process by biomedical animator Drew Berry can be found on YouTube <http://www.youtube.com/watch?v=OjPcT1uUZiE>.

[†] There is a distinction between the protein complex and the site at which the complex is loaded, though both are sometimes referred to as origins. In this thesis, initiation relates to the complex, while replication (or passive replication) relates to the site.

We first review licensing, the spatial aspect of the replication program, in different organisms. The best-understood eukaryote is *Saccharomyces cerevisiae* (budding yeast). In this single-cell organism, licensing occurs at the ACS (ARS consensus sequence), an 11-basepair (bp) sub-sequence in the roughly 100-bp domain known as the autonomously replicating sequence (ARS) [5]. Because licensing is sequence specific, the origin positions are well defined and are considered to be deterministic. In contrast, the licensing in budding yeast's distant relative *Schizosaccharomyces pombe* (fission yeast) is not determined by any particular sequence but takes place in kilo-basepair-sized (kb-sized), AT-rich, intergenic regions [13]. The licensing in this case may be considered stochastic, as an origin can lie anywhere within these kb-sized regions with high probability and probably within other regions also with lower probability. For multicellular organisms, genome-wide studies in *Drosophila melanogaster* (fruit flies) suggest that licensing occurs in domains that are roughly 10 kb in size and that these domains have complex sequence features that relate to open chromatin structure [14, 15]. Analyses of human replication experiment show that licensing occurs in roughly 100-kb zones that correspond to gene-rich and transcriptionally active regions [16, 17]. The most stochastic case is found in the embryos of *Xenopus laevis* (African clawed frog), where licensing seems possible everywhere along the genome [18]. A more recent study shows that licensing is uniform only up to mega-basepair (Mb) domains [19].

The temporal aspect of the replication program is more confusing. Bulk or population experiments probing replication show that some parts of the genome replicate, on average, earlier than others, forming a replication timing pattern [14, 16, 20]. A common interpretation of this result is that initiations are temporally ordered to ensure controlled duplication. An origin is thus often classified as early or late in the literature. This description implies a roughly deterministic view that a particular origin on a particular site (or in a particular zone) would initiate at a particular time in every cell. However, single-molecule experiments have shown that the initiation of origins is stochastic; i.e., a nominally late origin can fire early with non-negligible probability [21, 22]. In this thesis, we focus on budding yeast and frog embryos, which are at the two extremes of the range of licensing behaviour. In exploring the timing behaviour in these two organisms, we hope to 1) clarify the stochastic nature of the temporal aspect of replication in all eukaryotes and 2) elucidate the yet unknown mechanism for controlling origin-initiation and genome-duplication time with stochastic initiators.

1.2 Overview of modelling

Genome duplication is a complex process that can be modelled at different scales. On the molecular level, most models concern the structural mechanism and biochemical steps involved in the function of replicative machinery. These steps form a molecular network that can be analyzed on a systems level. Recently, Brummer *et al.* modelled the major molecular events using a system of coupled ordinary differential equations. Using this systems approach, they demonstrated how the interaction network can lead to re-replication when the activator CDK is upregulated too early in S phase [23]. Our modelling is on a coarser scale: we group all the molecular steps involved in replication into three kinetic elements: the position and time of initiation (or the replication program), the fork velocity, and the position and time of coalescences. As we will show in this thesis, such a coarse-grained approach connects well to large-scale experimental data and provides a good description of the replication kinetics in eukaryotes.

Modelling the replication process using simple kinetic processes is not a new idea. An early effort by Bertuzzi *et al.* presented a deterministic model where every cell replicates at the same rate [24]. Their model has been recently used to extract the fraction of cells in S phase at a given time and the length of S phase from flow-cytometry data [25]. A more sophisticated approach where the licensing and initiation are modelled as Poisson processes was developed by Cowan [26]. Using the Poisson model, he investigated many aspects of replication, such as the genome duplication time, the number of origins used in each S phase (not all potential origins initiate because of passive replication), and the statistics of Okizaki fragments. Though rich in content, Cowan's work was only marginally connected with experiments. In this thesis, we develop similar but more-general models and focus on their application to recent experimental data.

Our modelling started with the realization that the DNA replication process is formally equivalent to a one-dimensional version of the stochastic nucleation-and-growth model introduced in statistical physics to describe first-order phase transitions (Fig. 1.2). This model, often referred to as the Kolmogorov-Johnson-Mehl-Avrami (KJMA) theory of phase-change kinetics [27, 28, 29, 30, 31, 32], captures three aspects of phase trans-

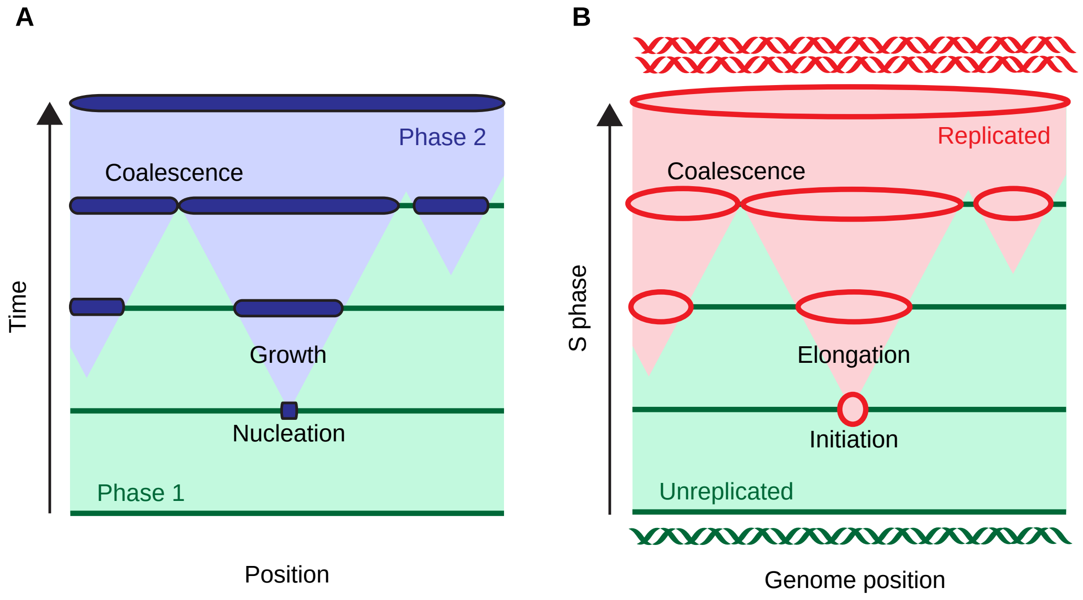


Figure 1.2: Schematic analogy between a nucleation-and-growth model and DNA replication. **A**. A one-dimensional nucleation-and-growth model that describes the phase transformation kinetics from Phase 1 to Phase 2. **B**. The DNA replication process. Darker lines correspond to unreplicated DNA; lighter bubbles correspond to replicated DNA.

formation: nucleation of the transformed phase, growth of the nucleated domains, and coalescence of impinging domains. Making a formal analogy between phase transformations and DNA replication, we map the kinetics of the DNA replication onto a one-dimensional KJMA model with three corresponding elements: initiation of potential origins, growth of replicated domains, and coalescence of replicated domains. The use of a two-state phase-transformation model implicitly incorporates the observation that, ordinarily, re-replication is prevented. That is, only two states (replicated and unreplicated) are allowed. Re-replication, which can occur in cancer cells, leads to additional states and is thus not considered in the present model.

Bechhoefer and Jun extended the KJMA model to include arbitrary temporal variations in the initiation rate and applied it to molecular-combing data from frog embryos to extract quantities such as the rate of origin initiation [33, 34, 35]. Such information has led to an appreciation of the role of stochastic effects in initiation [36], to models highlighting searching and binding kinetics in initiation timing [37], and to suggestions that initiation patterns may be universal across species [38]. In this thesis, we extend the formalism further to include arbitrary spatial distributions of initiation, providing a flexible formalism for describing eukaryotic replication.

The formalism that we will develop here is less biased than popular empirical analyses [20, 39], as it accounts for the effects of stochasticity and passive replication. Because it is analytic, it is also faster than simulation-based models [40, 41, 42]. Most importantly, it allows us to construct models that capture the complete replication kinetics quantitatively from experimental replication profiles. This ability makes possible a more complete understanding of the replication kinetics in budding yeast and frog embryos and leads to insights about the open question of initiation timing. While the licensing properties of the two organisms, as mentioned above, are at the two extremes of the deterministic-to-random spectrum, their initiation properties are very similar. The mechanism that seems to underlie the initiation of individual origins in budding yeast also ensures timely genome duplication in frog embryos. We will develop this theme in more detail throughout the thesis.

In the bigger picture, we hope that our work can eventually contribute to medical advances in replication-related diseases such as cancer. A hallmark of cancer is genome instability, which includes mutations and gross abnormality in chromosome structures [43]. Most tumourgenesis models attribute the cause of cancer to mutation and hyperactivity in genes known as oncogenes [2, 44]. These oncogenes are associated with kinetic elements such as re-replication [2] and stalled forks [44], which can cause DNA breakage and lead to gross chromosome structures [45]. Since the development of cancer relates to abnormal kinetic elements, one can incorporate these into the model. A recent work that included fork stalling suggests that the density of stalled forks could be an indicator of whether a cell is normal or cancerous [46]. That work illustrates the usefulness of quantitative modelling in understanding cancer. More elaborate and careful studies using such models can lead to quantitative comparisons between replication in cancer and normal cells and may reveal novel and specific targets for diagnosis and treatment.

1.3 Overview of experimental techniques

Our main motivation in developing new theory is to extract quantitative information from experiments in a rigorous and logical way. Among the many experiments that probe different aspects of replication (e.g., the genome-wide distribution of ORC [15, 47] and the spatiotemporal distribution of forks [48]), we focus on those that measure the variation of DNA content during S phase. These DNA measurements are usually easier than measurements of protein occupancy because DNA is much more abundant. Mathematically, we describe the variation of DNA content in S phase by the replication fraction $f(x,t)$, which is defined, at a population level, as the fraction of cells in a culture that has genome position x replicated at time t after the start of S phase. At the single-cell level, $f(x,t)$ can be interpreted as the cumulative probability of having the site x replicated at time t after the start of S phase.

How does the replication fraction $f(x,t)$ connect to our kinetic picture? As shown in Fig. 1.2, the position and timing of initiation plus the growth velocity of replicated domains determine the progress of replication. Our picture is that although each realization of genome duplication in a cell is stochastic [21, 22], all cells in the population follow the same underlying replication program. We characterize this program by a spatiotemporal initiation rate $I(x,t)$ (defined later in Sec. 2.1.1) and a fork velocity v, which is well approximated as a constant in several organisms (also discussed in Sec. 2.1.1). Thus, the replication fraction $f(x,t)$ is a functional of $I(x,t)$ and v. Below, we present some of the recent experimental techniques that probe the replication fraction. In general, the techniques can be labelled with four properties:

> **Genome mappability**. We define "mappable" to mean that the measurement can be mapped to a specific genome position x. Such measurements provide spatial information about $f(x,t)$. Methods of mapping include fluorescence in situ hybridization (FISH), complementary base pairing (used in microarrays), and sequence alignment (used in deep sequencing).
>
> **Temporal synchronization**. We define "synchronized" to mean that the measurement can be associated with a particular time in S phase. Such measurements provide temporal information about $f(x,t)$, where t is the time relative to the start of S phase. In practice, this involves synchronizing a population of cells so that one can release all cells in the culture to enter S phase at the same lab time (Fig. 1.3A, left plot). In doing so, one can assign a well-defined time in the cell cycle to a measurement. Such an assignment is not possible for an asynchronous cell culture, as the cells are

distributed throughout the cell cycle time at any lab time (Fig. 1.3A, right plot). Synchronization methods for yeasts include the use of alpha-factor, temperature-sensitive strands, and elutriation [49]. Multicellular organisms are generally more difficult to synchronize.

Single molecule. We define “single molecule” to mean that the measurement is from a single DNA fibre (Fig. 1.3B). Such measurements can provide different statistics of the replication fraction, such as its mean, variance, and covariance, whereas the bulk/population measurements provide only the mean. In this thesis, we focus on $f(x, t)$, which is the mean of the replication fraction. If larger single-molecule datasets were available, one could extend the analysis in the thesis to incorporate additional statistics. Older single-molecule techniques include electron micrography [50] and DNA fibre autoradiography [51]; more modern methods include DNA combing, which uses fluorography. As the names suggest, autoradiography involves labelling the DNA with radioactive substances, while fluorography involves labelling the DNA with fluorophores. The autoradiography assay is time consuming and has been replaced by the more efficient and economical fluorography.

Spatial coverage. We define “spatial coverage” to mean the domain size of x in $f(x, t)$ covered by the dataset. Single-molecule techniques often provide megabasepair (Mb) coverage. For organisms with short genomes (≈ 10 Mb) such as budding yeast, this can mean chromosome-wide coverage. Microarray and sequencing techniques can provide genome-wide coverage, regardless of the size of the genome. The ability to cover large parts of the genome is a major technological advance, as this allows a global comparison of origin properties. By contrast, the older, time-consuming techniques such as two-dimensional gel electrophoresis, which tests definitively whether a particular sequence can function as an origin, are suitable for probing only a handful of genome locations [52].

The above properties indicate the type of information that the experimental techniques can provide. Table 1.1 lists the techniques that are most relevant to our studies. In discussing the techniques’ advantages and limitations below, we hope to give a general view of the technological status of the field.

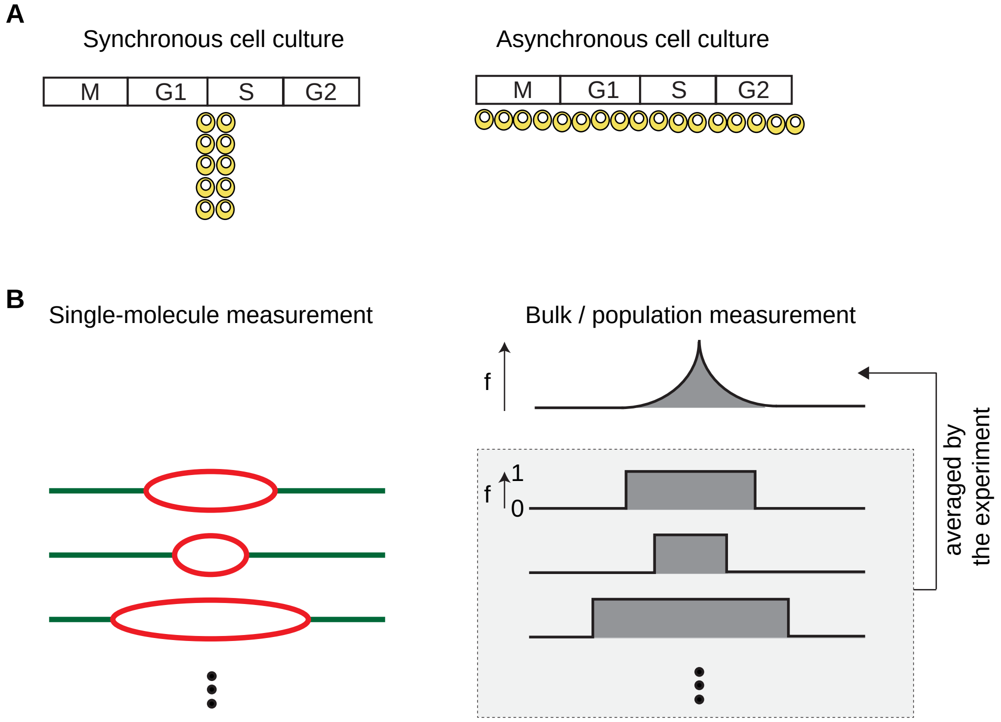

A
Synchronous cell culture
M
G1
S
G2
Asynchronous cell culture
M
G1
S
G2
B
Single-molecule measurement
Bulk / population measurement
f
f
1
0
averaged by
the experiment

Figure 1.3: A schematic diagram for the "synchronization" and "single molecule" properties of experimental techniques. **A**. Synchronous vs. asynchronous cell culture. **B**. Single molecule vs. bulk measurements. On the left, lines correspond to unreplicated DNA; bubbles correspond to replicated DNA, as in Fig. 1.2B. On the right, the unreplicated parts have replication fraction $f = 0$, while the replicated parts have $f = 1$. A bulk/population measurement is ideally equivalent to an average of the single-molecule replication fractions.

Technique	Map	Sync	Single Mole.	Coverage
FACS [53]	×	×✓	×	Genome
DNA combing [54]	×	✓	✓	Genome
FISH-combing [55, 21, 22]	✓	×✓	✓	Mb
Time-course microarray [20, 56, 57]	✓	✓	×	Genome
FACS-microarray/seq [16, 53, 58]	✓	×	×	Genome

Table 1.1: Examples of replication experiments and their properties. The symbol "×✓" denotes that either outcomes is possible.

Fluorescence-activated cell sorting (FACS). This technique probes DNA content by flowing cells through a channel one by one and measuring the fluorescence intensity from the cell's stained nuclear DNA (Fig. 1.4A) [59]. The detected intensity is proportional to the amount of DNA in the cell. The method is typically applied to an asynchronous cell culture in which a cell's DNA content varies between 1 copy and 2 copies of DNA, depending on its "position" in the cell cycle (Fig. 1.1). By sorting the cells with respect to their intensity, one can select cells that are in a particular phase of the cell cycle (Fig. 1.4B).

The intensity histogram, shown schematically in Fig. 1.4B, can be used to construct the replication fraction $f(t)$ and initiation rate $I(t)$ as a function of time. (The histogram provides no spatial information because the intensity is summed over the cell's genome.) An analysis of FACS histograms is provided in Chapter 5. The main advantage of this technique is that it is easy, fast and economical. Generally, the setup of flowing and counting cells is known as flow cytometry*.

DNA combing. In this technique, one labels DNA replicated before a predetermined time point with modified nucleotides, stretches out the DNA fibres in a controlled way onto a substrate, and detects the replication patterns with fluorescent antibodies (Fig. 1.5) [36, 60, 61]. Figure 1.5B shows a typical result. The fibres are "snapshots" that show replicated and unreplicated domains of the replicating DNA at t_p. The snapshots can be used to infer the initiation rate and the fork velocity [34, 35, 54]. As an illustration, from the combed DNA fibres, Herrick *et al.* compiled statistics for the size of replicated domains, the size of unreplicated domains, and the distance between centres of replicated domains as a function of time [54]. Since these fibres are sampled from across the genome, the statistics can be used to infer the fork velocity v and the genome-averaged initiation rate $I(t)$ as a function of time [34, 35, 54]. The inferred $I(t)$ and v for frog embryos in [54] are important quantities in the analysis presented in Chapter 6. Note that the combed fibres are not mapped to particular positions along the genome in this technique.

Advantages of DNA combing include low noise and the possibility to find correlations among initiations (e.g., [62]). The main disadvantage is that the size of combed fragments is small (100 kb to 1 Mb) relative to the typical size of mammalian genomes ($\mathcal{O}(10^2)$–$\mathcal{O}(10^3)$

*Throughout the thesis, we do not distinguish between "FACS" and "flow cytometry." Strictly speaking, the histogram in Fig. 1.4B is a flow cytometric histogram and not a FACS histogram, since only the flow but not the sorting is needed to generate the histogram. The sorting, however, is crucial for the FACS-microarray mentioned below; thus, "FACS-microarray" is the correct name there.

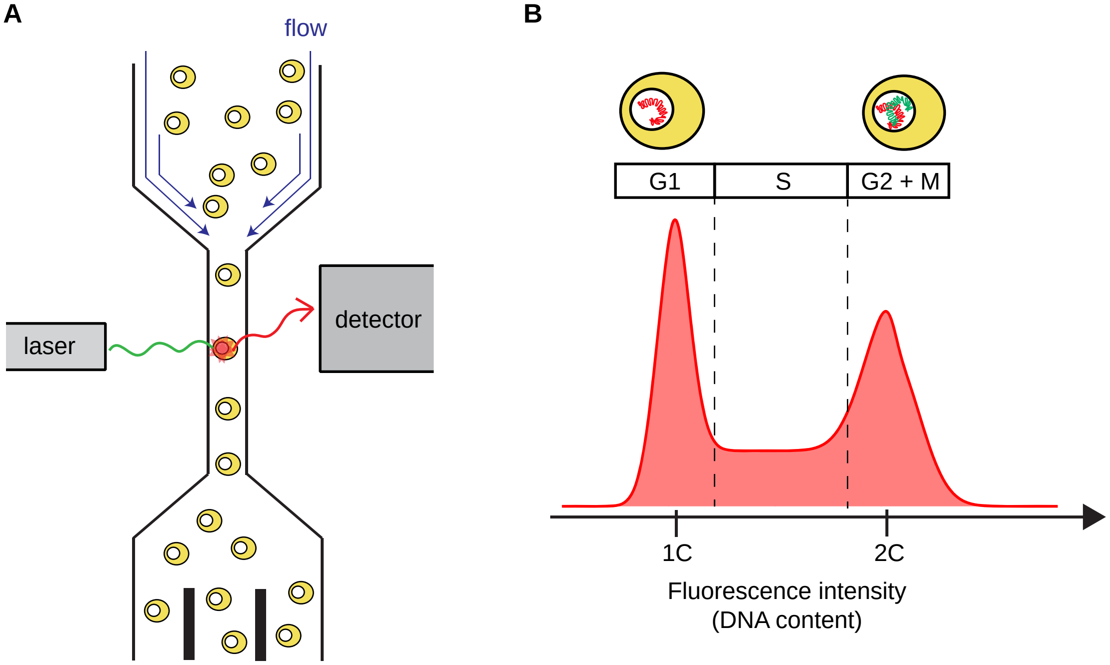


Figure 1.4: A schematic diagram of fluorescence-activated cell sorting (FACS). **A**. The setup. Cells are flowed one by one through a narrow channel. A laser beam excites the fluorophores incorporated into the cell's DNA. The detector gathers the fluorescence signals. Cells are flowed to different storage chambers depending on the detected signal. **B**. Histogram of detected fluorescence signal. The fluorescence intensity is proportional the cells' DNA content, which ranges from 1 copy (1C) to 2 copies (2C) of DNA.

Mb). The limitation in fibre size complicates analysis [63] and does not give information about large-scale spatial variations of the replication kinetics. Another disadvantage is the lack, to date, of an effective automated system to identify fibres. Due to this shortage, fibres have been studied mostly manually under the microscope, and only a relatively small fraction of all available fibres are used.

FISH-combing. In this technique, the combed DNA fibres are mapped to specific regions along the genome using FISH [55, 21, 22]. In contrast to DNA combing, this mappable technique provides position-specific information on the replicated and unreplicated domains, whose boundaries correspond to replication forks. This additional feature has allowed exploration of timing correlations between pairs of origin initiations [22] and motivated an analysis based on a reformulation of the KJMA model in terms of forks [17].

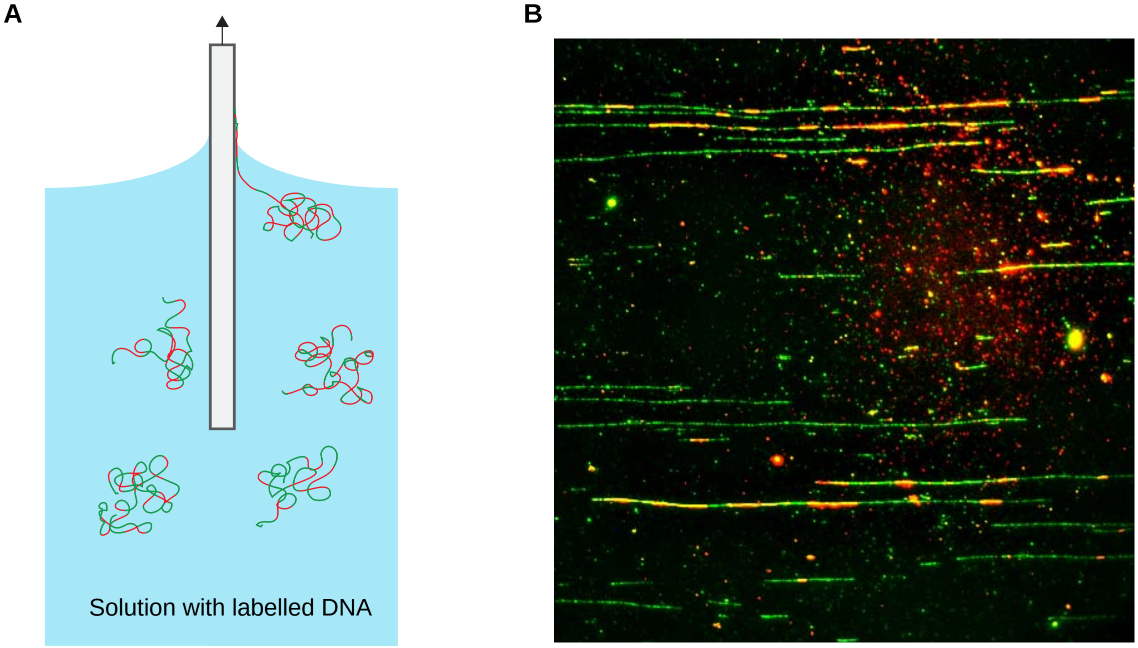


Figure 1.5: A schematic diagram of DNA combing. **A**. The combing processes. A glass slide coated with silane is dipped into a solution with labelled DNA. DNA molecules bind to the glass surface by their ends (only one is shown for clarity). When the glass slide is pulled up from the solution, the liquid and air interface "combs" (stretches and flattens) the DNA onto the glass surface. **B**. (Coloured in electronic file) Typical combed DNA fibres under the microscope. Regions of DNA that are replicated before a predetermined time t_p relative to the start of S phase are labelled with BrdU (a modified nucleotide) that is visualized with red antibodies. All DNA are also labelled with green anti-ssDNA antibodies. (Red and green merge to show yellow.) Courtesy of Dr. Nicholas Rhind.

The disadvantage of FISH-combing is that it has limited spatial coverage (see Table 1.1) and provides many fewer fibres for analysis.

Time-course microarray. Figure 1.6A shows a schematic diagram of the technique. The adjective "time-course" indicates that the probed cell cultures are synchronized. The DNA of a synchronized culture is extracted and hybridized onto a microarray chip. With synchronization, each culture can be assigned a time point t_p with respect to the start of S phase. The resulting measurement is a spatially resolved replication fraction $f(x, t = t_p)$ at time t_p. Repeating the procedure at multiple time points, one obtains the replication fraction $f(x,t)$ sampled with finite spatial and temporal resolution.

Compared to molecular combing, the main disadvantages of microarrays are the loss of information about cell-to-cell variability and the need for complicated data processing to remove artifacts. Fortunately, the second disadvantage can now be largely overcome by sequencing techniques [64]. The main advantages of both microarray and sequencing are the techniques' high-throughput and complete coverage of the genome at high resolution (kb – 100 bp). The availability of high-resolution, genome-wide data allows one to quantify in detail the complete spatiotemporal replication program, averaged over a population of cells. The analysis of a time-course-microarray dataset is presented in Chapter 3.

FACS-microarray. In this technique, cells in S phase are first separated from the asynchronous culture by FACS. Their DNA is then extracted and hybridized onto a microarray chip (Fig. 1.6B). The resulting measurement is a temporally averaged replication fraction $f(x)$. The analysis of data from this technique is the focus of Chapter 5. Compared to time-course microarrays, this technique averages out temporal features of the replication program; however, it is much more accessible because synchronization, which is difficult for most eukaryotes, is not needed. In other words, FACS-microarray is preferred for probing a wide range of organisms' spatiotemporal replication programs, while time-course microarray is better for understanding model organisms in great detail.

In addition to the techniques mentioned, other important techniques for understanding replication probe replication in real time and *in vivo*[*]. An example of the former is an *in vitro* experiment that follows the replication of stretched DNA fibres in real time [65]. This technique can be likened to a video version of DNA combing and has allowed direct measurement of replication fork movements [66, 67]. Since the observations are in real time, they are naturally "synchronized" to the lab clock. The main disadvantage is of course that such experiments cannot be done *in vivo*.

A popular technique for probing replication *in vivo* is microscopy. Viewing fluorescently labelled DNA and proteins in the nucleus can reveal the relation between replication and the active nuclear environment [68]. Microscopy studies suggest that replication forks are not randomly distributed but are clustered together in three-dimensional space to

[*] The Latin phrase "*in vivo*" means "in something alive" and describes measurements that probe replication in the nucleus. The Latin phrase "*in vitro*" means "in glass" and describes measurements made outside the cell in an artificial setup. Sometimes, the phrase "*in silico*" is used for computer simulations of such measurements.

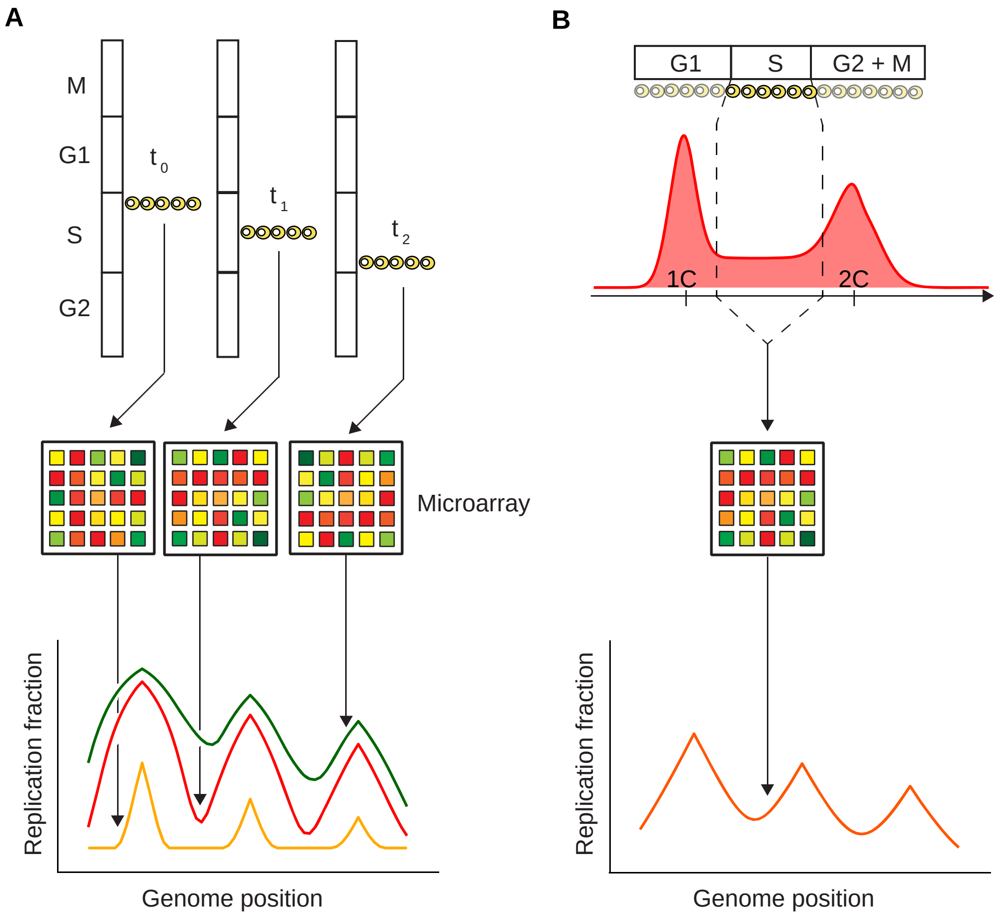


Figure 1.6: A schematic diagram of time-course microarray and FACS-microarray. **A**. Time-course microarray. The DNA from synchronous cell cultures are hybridized onto microarray chips or sequenced. The time labels ($t_0 < t_1 < t_2$) show how long the cell culture has been in S phase. Ideal results without instrumental noise are shown at the bottom. **B**. FACS-microarray. Cells progressing through S phase in an asynchronous cell culture are selected by FACS. The DNA are extracted from the cells, fragmented, and hybridized onto a microarray chip. Ideal results without instrumental noise are shown at the bottom.

form “replication factories” that replicate the DNA [69, 70]. Microscopy studies also show that DNA at the periphery of the nucleus are replicated later than DNA near the centre [71]. The major drawbacks of the technique are its inability to resolve individual DNA fibres and identify their precise genome positions. Recently, super-resolution microscopy techniques have provided microscopy images of replicating chromosomes with resolution (≈ 10 nm) that is an order of magnitude better than normal light microscopy (≈ 200 nm) [72]. Images using such techniques have resolved the individual components of replication factories. Coupling the analysis of replication kinetics to the nuclear environment is an exciting future direction.

1.4 Structure of the thesis

This thesis is organized as follows (the citations indicate that parts of the analyses presented are published in the corresponding references): We first present the theoretical framework in Chapter 2 [73, 74]. We then apply the theory to time-course microarray data and extract parameters that quantify the genome-wide spatiotemporal program of budding yeast in Chapter 3 [74]. Based on the parameters extracted, we propose and test a model for the control of origin initiation timing in yeast in Chapter 4 [74]. In Chapter 5, we seek to reconstruct the replication program from the more-accessible FACS-microarray experiments. In Chapter 6, we address the “random-completion problem,” which asks how replication-completion time can be controlled with stochastic origin licensing and initiation [73]. Lastly, we present concluding remarks and possible future directions in Chapter 7. There are three themes throughout the thesis: 1) the development of a general framework that can describe the replication programs of eukaryotes, 2) the application of models in the framework to various experiments, and 3) the advance in our understanding concerning the temporal aspects of replication control. Each chapter begins with an overview that connects to these themes.

Chapter 2

Theory

As discussed in Chapter 1, our picture of replication kinetics has three elements: the initiation of licensed replication origins, the growth of replication domains via replication forks, and the coalescence of domains (Fig. 2.2). For origin initiation, the theory represents the position and timing probabilistically and is sufficiently general to describe both deterministic and stochastic replication-kinetics scenarios. We assume only that initiation events are not correlated (discussed below in Sec. 2.1.1). For fork movements, we focus on the simplest case of a constant, deterministic velocity (discussed in Sec. 2.1.1). With these two elements, we derive formula for replication fractions and characterize origin properties. These form the basis of the models used in Chapter 3 and 4 to extract information from experiments on budding yeast and the inversion method presented in Chapter 5.

The third element, coalescence, is useful for deriving the distribution of replication-completion times. (The last coalescence defines the completion time.) We focus on a scenario where initiation is spatially uniform and derive explicit formula for the end-time distribution given power-law initiation rates. The results are used in Chapter 6 to address the random-completion problem in frog embryos. We also include a brief description of our simulation methods at the end of the chapter. In terms of the three themes mentioned in Sec. 1.4, this chapter presents a general framework that can describe the replication kinetics in eukaryotes. Sections 2.1 and 2.2 are based on [74] and [73], respectively.

2.1 Formalism for eukaryotic replication kinetics

2.1.1 Replication fraction

From our picture of the replication process, an organism's replication kinetics can be fully determined by the properties of the origins, which are specified through their licensed positions and ability to initiate throughout S phase, and the properties of the fork progression. While fork progression can in general be described as a spatiotemporal function $v(x, t)$, we assume, for the derivation below, the simplest case that all forks travel at a constant velocity v. This assumption is reasonable, as fork velocity has been observed in budding yeast and human cell lines to be roughly constant [48, 75]. The generalization to variable fork velocity is presented briefly in Appendix 3.A.2 and in more detail in [17, 76].

The neglect of stochasticity in fork movements, interpreted as using an effective, averaged fork velocity, is a reasonable approximation. Using a single-molecule approach, Lee *et al.* showed that bacteriophage fork progression, though marked by transient pauses that last for seconds, is essentially constant on the kb/min scale (Fig. 4 in [66]). Using the same approach, Tanner *el al.* showed that bacteria fork progression is also constant on the kb/min scale (Fig. 2 in [67]). In eukaryotes, fork-to-fork variation can result from forks that stall to repair misreplicated and damaged DNA [6, 77]. Gauthier and Bechhoefer modelled the effect of stochastic fork stalls on replication kinetics and concluded that the effective fork velocity is not affected until the stall density exceeds a threshold that is well above what is observed in normal replication [46]. Based on these findings, we argue that an averaged fork velocity is reasonable for the scale relevant to our modelling[*].

Following previous work [33], we describe the kinetic properties of origin via an initiation rate $I(x, t)$, defined as the number of initiations per time per unreplicated length of DNA, at genome position x and at time t after the start of S phase. Such a description implies that origin initiations are not correlated. Even though there is evidence for correlation in some organisms (e.g., in frog embryos [62]), the correlation has only a minor effect on the replication kinetics. In simpler organisms such as budding yeast, correlation among

[*] Using single-molecule "magnetic tweezers" experiments, Danilowicz *et al.* showed that the unzipping of double-stranded DNA under constant force is not continuous but consists of stochastic jumps and pauses that are sequence dependent [78]. Comparing this result to the much more constant velocities observed in replication experiments *in vitro* [66, 67], we speculate that the unzipping of DNA by helicases is actively regulated to proceed at a roughly constant rate.

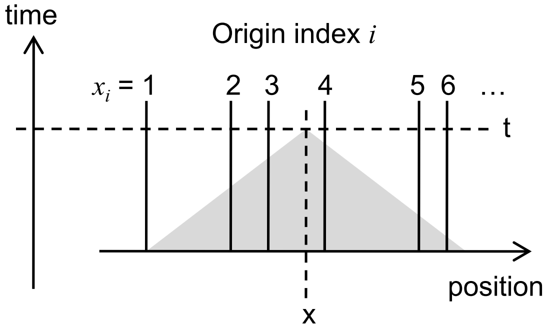


Figure 2.1: Illustration of Kolmogorov's argument. The replication fraction at (x, t) equals to "1 − probability that no origins initiated within the shaded triangle △." Vertical lines indicate the positions of discrete origins; initiation can occur only along these lines.

initiations is insignificant [22].

Given $I(x,t)$, we can calculate the replication fraction $f(x,t)$ of site x on the genome at a time t after the start of S phase [27, 33]. The result is

$$f(x,t) = 1 - \prod_{\triangle} \left[1 - I(x',t')\Delta x'\Delta t'\right], \tag{2.1}$$

where the product is over all area elements of $\Delta x \Delta t$ in the shaded triangle △ depicted in Fig. 2.1 [27]. We note that the product in Eq. 2.1 is the key concept in the theory and will reappear in many of the quantities derived. In words, the replication fraction at a specific position and time is equal to one minus the probability that the position has not been replicated before that time. In the limit $\Delta x \to 0$ and $\Delta t \to 0$, one finds

$$f(x,t) = 1 - e^{-\iint_{\triangle} I(x',t')dx'dt'} . \tag{2.2}$$

Equations 2.1 and 2.2 are remarkable because they provide an essentially exact solution to a nontrivial many-body problem in $1+1$ (space and time) dimensions.

We start with a description of budding yeast, where origin positions are well defined and sequence specific. Denoting these positions by x_i, we can describe an origin by its initiation rate $I_i(x,t) = \delta(x - x_i)I_0(x,t)$, where $\delta(x)$ is the Dirac delta function, which is zero for all $x \neq x_i$. The total rate $I(x,t)$ is the sum of all $I_i(x,t)$. To make the double integral in Eq. 2.2 explicit, we define a local measure of origin firing,

$$g(\Delta x_p, t) = \int_{x_p}^{x_{p+1}} \delta(x - x_i)dx \int_0^t I_0(x, t')dt', \tag{2.3}$$

for the interval $[x_p, x_{p+1})$, where the index p comes from discretizing a genome of length L:

$$\Delta x = \frac{L}{N}, \quad x_p = p(\Delta x), \quad p = 0, 1, ..., N - 1. \tag{2.4}$$

The function $g(\Delta x_p, t) = 0$ unless there is an origin contained in the interval $[x_p, x_{p+1})$. Replacing the double integral in Eq. 2.2 by a sum using Eq. 2.3, we obtain

$$f(x, t) = 1 - \exp\left[-\sum_{p=0}^{N-1} g\left(\Delta x_p,\, t - \frac{|x - x_p|}{v}\right)\right], \tag{2.5}$$

where v is the fork velocity. The Δx_p in the first argument of $g(x, t)$ is an interval, while the x_p in the second argument is a point. The second argument, $t - |x - x_p|/v$, is the time along the edge of the triangle in Fig. 2.1. Biologically relevant $g(x, t)$ should satisfy the following constraints:

1. $g(\Delta x_p, t < 0) = 0$. This means that the initiation rate is zero $[I(x, t < 0) = 0]$ before the start of replication. Applying this constraint to Eq. 2.5, we see that the sum is limited to the domain $(x - t/v) \leq x \leq (x + t/v)$.
2. $\frac{d}{dt} g(\Delta x_p, t) \geq 0$. Since $g(\Delta x_p, t) = \int_x^{x+\Delta x} \int_0^t I(x', t')dt'dx'$, this is equivalent to $I(x, t) \geq 0$, meaning that the initiation rate cannot be negative.
3. $g(\Delta x_p, t) \geq 0$. This is a direct consequence of constraints 1 and 2.

One can generalize Eqs. 2.3 and 2.5 to a continuous version where origins can be anywhere; i.e., the x_i that defines the budding yeast origin is now a continuous variable instead of a finite set of positions. Renaming it x', we rewrite Eq. 2.3 as

$$g(x', t) = \int_0^t I(x', t')dt' \tag{2.6}$$

and Eq. 2.5 as

$$f(x,t) = 1 - \exp\left[-\int_0^L g\left(x',\, t - \frac{|x-x'|}{v}\right) dx'\right]. \tag{2.7}$$

One can now choose $I(x,t)$ to appropriately describe any organism. For example, if the licensing of an origin occurs in a particular zone along the genome, one might describe this origin with $I_i(x,t) = N(x_i, \sigma_i) I_i(t)$, where $N(\mu, \sigma)$ is the normal distribution with mean μ and standard deviation σ. The genome-wide $I(x,t)$ is the sum of all origin contributions.

In application to the bulk experiments mentioned in Sec. 1.3, we fit Eq. 2.5 to a time-course microarray dataset, where the budding yeast cell culture was synchronized to enter S phase at roughly the same time and where the replication fraction was measured for specific genome position [57]. The experiment generated the genome-wide $f(x,t)$ at eight time points, and we fit $f(x,t)$ in order to extract $I(x,t)$ and v, which fully characterizes the replication kinetics. The details are presented in Chapter 3.

For asynchronous and unmappable techniques such as fluorescence-activated cell sorting (FACS), the resulting histogram of DNA content is related to the genome-averaged replication fraction $f(t)$. In principle, one can average Eq. 2.5 over space and fit to such data to extract $I(x,t)$ and v. In practice, this data can, at most, provide information on the time dependence of the rate because the degeneracy in parameters after the genome-average is too large. For asynchronous but mappable techniques such as FACS-microarray, the resulting replication fraction profile is a time-averaged spatial profile $f(x)$. To analyze such data, one can average Eq. 2.5 over time and extract $I(x,t)$ and v. In Chapter 5, we present the the analysis of both FACS and FACS-microarray data. In particular, we will discuss methods to directly invert $I(x,t)$ from averaged $f(x)$ and $f(t)$.

Recently, by considering the problem in a light-cone coordinate, Baker derived an inversion formula that, given v, directly transforms $f(x,t)$ to $I(x,t)$ [76]:

$$I(x,t) = \frac{v}{2}\left(\frac{1}{v^2}\partial_{tt} - \partial_{xx}\right) \ln\left[1 - f(x,t)\right]. \tag{2.8}$$

Replacing the continuous double derivatives with their numerical counterparts and defining

$H(x,t) = -\ln[1 - f(x,t)]$, one obtains a discrete version

$$\begin{aligned} I(x_r, t_s) = & \frac{1}{2v}\left[H(x_r, t_{s+1}) - 2H(x_r, t_s) + H(x_r, t_{s-1})\right] \\ & - \frac{v}{2}\left[H(x_{r+1}, t_s) - 2H(x_r, t_s) + H(x_{r-1}, t_s)\right], \end{aligned} \tag{2.9}$$

where r and s are indices that run through the discretized genome position and time. The inversion is an advance because it is in general much faster than fitting and because the $I(x,t)$ extracted is model free and may better reflect the real replication program. However, direct application of Eq. 2.9 to data can be troublesome. First, experimental data are often noisy, and the double numerical derivatives amplify the noise. Second, the temporal resolution of the data is usually lower than the spatial resolution, and the formula does not take the variable resolution scales into account. We will comment on a strategy to deal with these issues in light of the reconstruction method discussed in Chapter 5.

2.1.2 Characterization of origins

In the biology literature, origins are often described as “early” or “late” and “efficient” or “inefficient” (e.g., [57]). These classes are good for qualitative discussion but are insufficient to fully characterize the origins. In our modelling, we quantitatively characterize the origins by firing-time distributions. The properties of the origins can be easily summarized as the statistics of the distribution. For instance, the first moment of the distribution shows whether an origin is early or late firing, and the second moment shows the precision of the firing time. In general, one should also assign a licensing distribution for each origin to describe its spatial properties. However, for simplicity and clarity, we focus on the budding-yeast scenario, where each origin has a specific location, i.e., a δ-function for the licensing distribution. One can use wider distributions to characterize origins in other organisms.

We start by stating the cumulative initiation probability $\Phi(x_p, t)$ in terms of $g(\Delta x_p, t)$ [79]:

$$\Phi(x_p, t) = 1 - e^{-g(\Delta x_p, t)}. \tag{2.10}$$

Since $\Phi(x_p, t)$ is non-zero only for $x_p \leq x_i \leq x_{p+1}$, we introduce

$$\phi_i(t) \equiv \frac{d}{dt}\Phi(x_i \in \Delta x_p, t) \equiv \frac{d}{dt}\Phi_i(t) \tag{2.11}$$

to denote the initiation probability density of each origin i. The x_p in Eq. 2.10 is related to the discretization, while the subscript i in Eq. 2.11 relates to the origin position. Rearranging Eq. 2.10 for $g(\Delta x_p, t)$ and differentiating, we obtain the relationship between the initiation rate and the firing-time distribution[*]:

$$I_i(t) \equiv I(x_i \in \Delta x_p, t) = \frac{\phi_i(t)}{1 - \Phi_i(t)}. \tag{2.12}$$

The above equation shows that the initiation rate is the probability that an origin initiates between t and $t + dt$, given that it has not initiated before time t.

Using the probability distributions in Eq. 2.11, we define potential efficiency p_i as the probability that origin i would fire by the end of S phase if there were no passive replication; i.e., $p_i \equiv \Phi_i(t = t_{\text{end}})$ with t_{end} being the typical length of S phase. By contrast, the efficiency of any origin decreases because of passive replication by neighbouring origins. The utilized efficiency of the j^{th} origin thus depends not only on $\phi_j(t)$ but also with the $\phi_k(t)$ of neighbouring origins. Mathematically, the utilized efficiency E_j of the j^{th} origin is

$$E_j \equiv \int_0^{t_{\text{end}}} dt \;\; \phi_j(t) \prod_{k \neq j} \left[1 - \Phi_k \left(t - \frac{|x_k - x_j|}{v} \right) \right], \tag{2.13}$$

where the integrand can be seen an effective initiation probability density that represents the probability that the j^{th} origin initiates between t and $t + dt$ and that all the other origins have not initiated to passively replicate the j^{th} origin before time t.

Recently, de Moura *et al.* proposed that the replication kinetics is largely affected by origin "competence," defined as the probability that an origin is licensed [42, 81]. The biological picture is that an origin may not be licensed every time in every cell (thus, the competence of an origin is $\leq$1). Mathematically, the competence q_i of origin i, defined as $q_i \equiv \Phi_i(t = \infty)$ [81] differs from the potential efficiency [$p_i \equiv \Phi_i(t = t_{\text{end}})$] only in the integration limit (∞ vs. t_{end}). In practice, one can hardly distinguish an origin with high

[*] $I_i(t)$ is also known as a hazard function [80].

q_i but low p_i from an origin with low q_i and p_i by investigating the replication fraction, as the former origin contributes to replication significantly only after t_{end}. For this reason, a model that uses origins with $q_i = 1$ but varying p_i is as effective as a model that uses origins with varying q_i and p_i but needs fewer parameters. In this thesis, we use the former model.

Passive replication implies an interesting phenomenon: a group of nearby origins forms an effectively earlier-firing and more-efficient origin. Since it is always the earliest initiation in the cluster that counts, the effective initiation probability of the cluster is the extreme (smallest) value distribution of all the $\phi_j(t)$ in the cluster. To distinguish between the origins in a cluster and the final effective origin, we call the former "initiators" and the latter "origins." Biologically, this correspond to the scenario that multiple initiators—the protein complexes that can initiate the unwinding and synthesis process—are loaded onto a small stretch of DNA. This picture and its biological implications are the focus of Chapter 4. Using the new terminology, for n initiators near x, the effective cumulative initiation probability of the origin is

$$\Phi_{\text{eff}}(x,t,n) = 1 - \prod_{j=1}^{n} \left[1 - \Phi_j \left(t - \frac{|x - x_j|}{v} \right) \right] . \tag{2.14}$$

For large n, extreme-value theory asserts that Eq. 2.14 tends toward the Weibull distribution (Eq. 4.1) for a large class of functions $\Phi_j(t)$ whose probability densities $\phi_j(t)$ are bounded from one side (0 for $t < 0$) [82]. We note that the product in Eq. 2.14, which captures the notion of "the first among many," is the same as that in Eqs. 2.13 and 2.1. As we will mention in the following section, the replication-completion time, being defined by the last coalescence, also relates to a product of cumulative probabilities. The appearance of the same mathematical form in the effective initiation time and replication-completion time shows that the two timings share the same underlying principle.

2.2 Replication-completion-time distribution

An important feature of any replication program is that it should replicate the DNA in a timely manner. The quantity that reflects this feature is the distribution of replication-

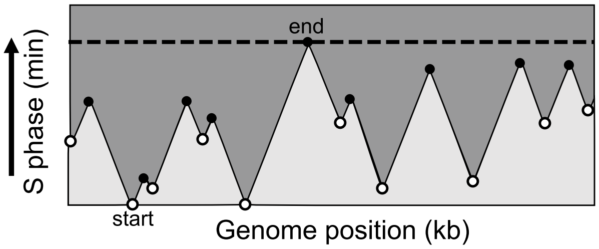


Figure 2.2: A schematic of the DNA replication model. A horizontal slice in the figure represents the state of the genome at a fixed time. The lighter (darker) grey represents unreplicated (replicated) regions. Open circles denote initiated origins, while filled circles denote coalescences. The dark dotted line cuts across the last coalescence, which marks the completion of replication. The slope of the lines connecting the adjacent open and filled circles gives the inverse of the fork velocity.

completion times (or, simply, the end-time distribution). In this section, we derive the end-time distribution for the case of embryonic cells and use it to address the "random-completion problem" in Chapter 6. At the time of our study, a uniform initiation rate across the genome was thought to be a good description for replication in frog embryos; however, recent work has revealed an apparent spatial inhomogeneity in the initiation [19]. We will present the theory for $I(x,t) = I(t)$, i.e., an initiation rate that varies in time but not space, and will comment on the effect of spatial inhomogeneity at the end of Sec. 2.2.3.

2.2.1 Distribution of coalescences

Our model, which assumes a constant fork velocity, results in a deterministic replication pattern for each realization of licensing and initiation. Figure 2.2 illustrates such deterministic replication and shows that, except at the boundaries, there is a one-to-one mapping between the initiations and the coalescences. It follows that every distribution of initiations $\phi_i(t)$ has an associated distribution of coalescences $\phi_c(t)$. Since the completion of replication is defined by the last coalescence, the problem of determining the time needed to replicate a genome of finite length is equivalent to that of determining the distribution of times at which the last coalescence occurs. This distribution is the end-time distribution $\phi_e(t)$.

The temporal program of stochastic initiation times is governed by an initiation rate

$I(t)$, defined as the rate of initiation per unreplicated length per time. In writing down the initiation rate as a simple function of time, we are implicitly averaging over any spatial variation and neglecting correlations in neighbouring initiations. Below, we derive an analytical approximation to the end-time distribution function for arbitrary $I(t)$. This analytical result will allow us to investigate how initiation programs affect the timing of replication completion.

Taking the limit Δx and $\Delta t \to 0$ in Eq. 2.1 for $I(x,t) = I(t)$, we see that for an infinitely long genome, the fraction f of the genome that has replicated at time t is given by [33]

$$f(t) = 1 - e^{-2vh(t)} \ , \tag{2.15}$$

where v is the fork velocity (assumed constant), $h(t) = \int_0^t g(t')dt'$ and $g(t) = \int_0^t I(t')dt'$*. Equation 2.15 predicts that an infinite time is needed to fully duplicate the genome; however, since all real genomes have finite length, they can be fully replicated in a finite amount of time. During the course of replication, as long as the number of replicated domains is much greater than one, the infinite-genome model is reasonably accurate. However, since the number of domains is small at the beginning and end of replication ($f \to 0$ and $f \to 1$), we expect discrepancies in those regimes. In particular, to calculate the finite replication time expected in a finite genome, we need to go beyond the calculation of replication fraction.

We begin by introducing the hole distribution, $n_h(x,t) = g^2(t)\exp[-g(t)x - 2vh(t)]$ which describes the number of "holes" of size x per unit length at time t [33]. A "hole" is the biologists' term for an unreplicated domain surrounded by replicated domains. Since a coalescence corresponds to a hole of zero length, we write the coalescence distribution as $\phi_c(t) \propto n_h(0,t)$. Normalizing by imposing the condition $\int_0^\infty \phi_c(t)dt = 1$, we find

$$\phi_c(t) = \frac{2vL}{N_o} g^2(t) e^{-2vh(t)} \ , \tag{2.16}$$

where L is the genome length and N_o the expected total number of initiations. Note that N_o is also the total number of coalescences because of the one-to-one mapping discussed

*Equation 2.15 can also be derived from Eq. 2.7 by 1) replacing $g(x', t - |x - x'|/v)$ by $g(t - |x - x'|/v)$, 2) noticing that the function is 0 outside $[x - vt, x + vt]$ and symmetric around x, and 3) making the change of variable $t' = t - |x - x'|/v$ and $vdt' = dx'$.

previously. One can calculate N_o via

$$N_o = L \int_0^\infty I(t)[1 - f(t)]dt = L \int_0^\infty I(t)e^{-2vh(t)}dt \ , \tag{2.17}$$

where the factor $[1 - f(t)]$ arises because initiations can occur only in unreplicated regions. The integrand in Eq. 2.17 divided by N_o is the initiation distribution $\phi_i(t)dt$, which corresponds to the number of initiations between time t and $t + dt$.

Given the initiation distribution, we picture the initiations as sampling N_o times from $\phi_i(t)$. This implies that N_o independent coalescence times are sampled from $\phi_c(t)$. The replication-completion time, finite on a finite genome, can then be associated with the largest value of the N_o coalescence times, and the end-time distribution is the distribution of these largest values obtained from multiple sets of sampling from $\phi_c(t)$. At this point, we apply extreme-value theory (EVT) to calculate the end-time distribution.

2.2.2 Digression on extreme-value theory

EVT is a well-established statistical theory for determining the distributional properties of the minimum and maximum values of a set of samples drawn from an underlying "parent" distribution [83, 82]. The properties of interest include the expected value, fluctuations around the mean, frequency of occurrence, etc. EVT plays a key role in the insurance industry: for example, the "100-year-flood" problem that asks for the expected maximum water level over 100 years is an extreme-value problem [84]. In physics, EVT has attracted increasing interest and has been applied to analyze crack avalanches in self-organized material [85], degree distribution in scale-free networks [86], and many other problems.

EVT is powerful because of its universality. The key theorem in EVT states that the distribution of the extremes of an independent and identically distributed random variable tends to one of three types of extreme value distributions, the Gumbel, Fréchet, and Weibull distributions, depending only on the shape of the tail of the underlying distribution*. The universality of the extreme value distribution with respect to the underlying distribution is similar to that of the better-known Central Limit Theorem [89]. For an underlying distribution with an unbounded tail that decays exponentially or faster, the distribution of the

*Apparently, the origin of extreme-value theory can be traced back to the work of Fréchet in 1927 [87] and of Fisher and Tippett in 1928 [88].

extremes tends to a Gumbel distribution. Such is the case of *Xenopus* embryos since the underlying distribution, the coalescence distribution $\phi_c(t)$, is approximately proportional to $e^{-\tau^4}$, where τ is a dimensionless time. The decay rate $e^{-\tau^4}$ follows from applying Eq. 2.16 to the observation that $I(t) \sim t^2$ in *Xenopus* embryos (Sec. 6.2). The other initiation rates we consider also lead to the Gumbel distribution.

The Gumbel distribution,

$$\rho(x) = \frac{1}{\beta}\exp\left(-x - e^{-x}\right), \qquad x = \frac{t - t^*}{\beta}, \tag{2.18}$$

depends on only two parameters, t^* and β [83, 90, 82]. The former is a "location" parameter that gives the mode of the distribution. The latter is a "scale" parameter proportional to the standard deviation. We follow standard procedures to obtain t^* and β as a function of the initiation rate and the fork velocity [83, 90]. The main step is to recognize that the cumulative end-time distribution $\Phi_e(t)$, which has a Gumbel form, is equal to the product of N_o cumulative coalescence distributions, each resulting from the same initiation distribution $\phi_i(t)$. In other words, the probability that N_o coalescences occur at or before time t is equivalent to the probability that the last of them occurred at or before time t, which is also the probability that the replication will finish at or before time t. For our case, we find that the mode t^* is determined implicitly by

$$N_o\left[1 - \Phi_c(t^*)\right] = 1 \tag{2.19}$$

and $\beta \approx 1/[N_o\phi_c(t^*)]$. In Eq. 2.19, $\Phi_c(t)$ is the cumulative distribution of $\phi_c(t)$; thus, $[1 - \Phi_c(t)]$ is the probability that a coalescence would occur at or after time t. Equation 2.19 then implies that given a total of N_o coalescences, t^* is the time after which the expected number of coalescences is one, and therefore, the typical end time. The Gumbel form of the end-time distribution is one of our main results, as it allows quantitative comparison between the fluctuations of completion times resulting from different initiation rates.

2.2.3 Replication-completion distribution for power-law initiation

Below, we derive the end-time distribution for a power-law initiation rate $I_n(t) = I_n t^n$ (where $n > -1$) and a delta-function initiation rate $I_\delta(t) = I_\delta \delta(t)$. In the power-law case, $h(t) \propto t^{n+2}$, while for the δ-function case, $h(t) \propto t$. From Eq. 2.16, both initiation forms give rise to coalescence distributions that decay exponentially or faster, and thus, both forms will lead to an end-time distribution of the Gumbel form. Using these initiation rates, we see that the coalescence distribution given by Eq. 2.16 is completely determined by three parameters: the fork velocity v, the initiation strength prefactor (I_n or I_δ), and the initiation form [n or $\delta(t)$]. The relationship between these three parameters and the two Gumbel parameters reveals how different "initiation strategies" affect the completion time.

We write the cumulative distribution $\Phi_c(t)$ of the coalescences as $1 - \int_t^\infty \phi_c(t')dt'$. Using integration by parts, we obtain

$$\int_t^\infty \phi_c(t')dt' = \frac{L}{N_o} g(t)e^{-2vh(t)} - \frac{L}{N_o}\int_t^\infty I(t')e^{-2vh(t')}dt' \ . \tag{2.20}$$

Substituting Eq. 2.20 into Eq. 2.19, we obtain a transcendental equation

$$2vh(t^*) = \ln\left[(1-\alpha)Lg(t^*)\right], \quad \alpha = \frac{\int_{t^*}^\infty I(t)e^{-2vh(t)}dt}{g(t^*)e^{-2vh(t^*)}} \tag{2.21}$$

that relates the initiation parameters to t^*. For the width, Eqs. 2.16 and 2.21 give

$$\beta = \frac{1-\alpha}{2vg(t^*)} \ , \tag{2.22}$$

indicating that the width of the end-time distribution β is inversely proportional to $g(t^*)$. Since $g(t)$ is the integral of $I(t)$, and since $LI(t)dt$ is the number of initiations in the given time interval, $Lg(t^*)$ is the number of origins that would have initiated during S phase if there were no passive replication. In other words, $g(t^*)$ is the lower bound on the average number of potential origins per length ("average" here is over an ensemble of genomes). It is the lower bound because the potential origins that would have fired with a longer S phase are not counted. At the end of Sec. 6.3.2, we compare the inferred *in-vivo* bound on potential origin density with the experimental estimate.

In practice, given experimentally observed quantities such as v, t^*, and L, we solve Eqs. 2.21 and 2.22 numerically to determine the initiation prefactor (I_δ or I_n) and the width for different initiation forms [$\delta(t)$ or t^n]. Nevertheless, an analytical approximation of Eqs. 2.21–2.22 is possible, as the factor α is often small. For instance, in the power-law

$I(t)$ case, introduce a function $\eta(t) = be^{-at}$ that decays more slowly than $\phi_i(t)$. Imposing $\eta(t^*) = \phi_i(t^*)$ so that $\eta(t) > \phi_i(t)$ for $t > t^*$, we find α to be at most $\mathcal{O}(10^{-2})$. Neglecting α, we obtain the analytical approximations

$$\begin{aligned} I_n &\approx \frac{(n+1)(n+2)}{2vt^{*n+2}} \ln\left[\frac{L(n+2)}{2vt^{*n+2}}\right] && (2.23) \\ \beta &\approx \frac{n+1}{2vI_n t^{*n+1}} && (2.24) \end{aligned}$$

that show the explicit relationship between the initiation parameters and the Gumbel parameters. In summary, given a realistic initiation rate $I(t)$ and fork velocity v, we have shown that the distribution of replication end times tends toward a Gumbel form. We have also shown how the replication parameters relate to the location and scale Gumbel parameters analytically.

Our discussion of the end-time distribution thus far was based on an initiation rate $I(t)$ that is uniform across the genome, an assumption that was thought to be true in frog embryos. Recently, Labit *et al.* showed using FISH-combing that the initiation rate is inhomogeneous, at least for the 5-Mb region of the frog genome that they investigated [19]. In a previous analysis of DNA combing data done on frog embryos, Herrick *et al.* extracted a spatially uniform initiation rate $I(t)$ and a "starting-time" distribution $\psi_a(\tau)$ defined as the probability that a cell in the probed cell culture enters S phase between lab time τ and $\tau + \delta\tau$ [54]. (Here, t is the time relative to the start of S phase, and τ is the lab time.) In using a spatially averaged $I(t)$, the assumption was that all analyzed DNA fibres started replication at $t = 0$. However, as the authors pointed out, this was likely not true for two reasons: 1) the synchronization of the cell culture was imperfect, and 2) different regions of the genome may start replication at different times. Thus, they introduced the starting-time distribution to summarize the combined effect of the imperfect synchrony and spatial inhomogeneity in the initiation rate.

Building on the idea that inhomogeneous initiation rate effectively leads to a starting-time distribution, we consider a simple scenario that illustrates how spatial variations in initiation rate affects the end-time distribution. Suppose that the genome can be separated into m equi-length regions that all have the same replication kinetics, except for the replication-starting time. That is, the initiation rate for a region r can be described by

$I(t-t_r)$. (Note $I(t<0)=0$.) Denoting the starting time of region r as t_r, the cumulative end-time distribution for the region can be described as

$$\Phi_r(t) = \exp\left(-e^{-\frac{t-t^*-t_r}{\beta}}\right) \,, \tag{2.25}$$

where t^* and β are the mode and width of the distribution, respectively. The cumulative end-time distribution for the entire genome is then

$$\Phi_e(t) = \prod_{r=1}^{m} \Phi_r(t) = \exp\left(-\sum_{r=1}^{m} e^{-\frac{t-t^*-t_r}{\beta}}\right) . \tag{2.26}$$

Assuming that m is large and the starting times t_r are normally distributed with standard deviation σ, Eq. 2.26 becomes

$$\begin{aligned}\Phi_e(t) &\approx \exp\left(-\frac{m}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-\frac{t-t^*-t_r}{\beta}} e^{-\frac{(t_r)^2}{2\sigma^2}} dt_r\right) \\ &= \exp\left(-m\exp\left(-\frac{t-t^*-\sigma^2/2\beta}{\beta}\right)\right) \\ &= \exp\left(-\exp\left(-\frac{t-t^*-\beta\ln(m)-\sigma^2/2\beta}{\beta}\right)\right) . \end{aligned} \tag{2.27}$$

For a spatially uniform initiation rate, $\sigma = 0$, and the mode and width of the genome-wide end-time distribution is $t^* + \beta\ln(m)$ and β, respectively. In other words, a spatially inhomogeneous initiation rate shifts the mode to a later time by $\sigma^2/2\beta$. We will apply this result to the study of frog embroys in Sec. 6.3.2. For general $I(x,t)$ and $v(x,t)$, Gauthier *et al.* derived an approximate end-time distribution based on rate equations of fork movements [17].

2.3 Simulation

In addition to the analytical framework of the replication kinetics presented above, we also perform simulations of the replication process. The simulations are useful in confirming the analytical results, in generating test data for fitting and inversion, and in addressing cases where analytical approach becomes cumbersome (e.g., Sec. 6.5). The simulation, as

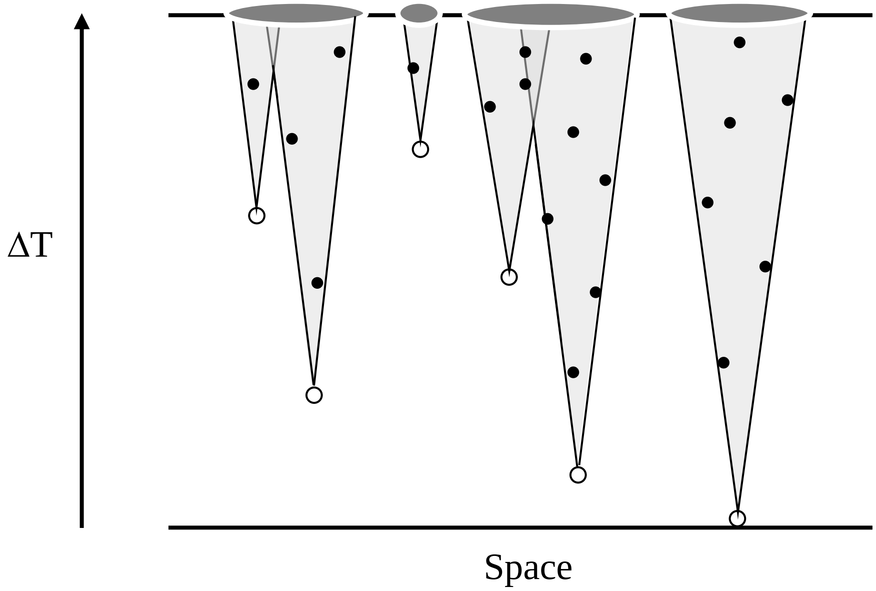


Figure 2.3: A schematic diagram of the phantom-nuclei algorithm. Open circles correspond to initiated origins. Dots correspond to passively replicated origins, or "phantom" nucleations, that are eliminated by the algorithm. The algorithm outputs the replicated (grey bubbles) and unreplicated (horizontal line) domain sizes. Figure reproduced with permission from Phys. Rev. E [Vol. 71, 011908, Fig. 7] [33] (Copyright ©2005, American Physical Society).

with our kinetic picture, is composed of several routines that correspond to the licensing and initiation of replication origins, the growth of replication domains via replication forks, and the coalescence of domains.

Licensing. For budding yeast, licensing amounts to assigning a number to the origin position. For frog embryos, the origin position is assigned to a uniform random number in [0,L], where L is the length of the simulated genome. In general, one can form the desired distribution by Monte Carlo methods [91].

Initiation. Given the form of the firing-time distribution or initiation rate, which we choose to be integrable and invertible, we use the Monte Carlo transform method to transform a uniform random number into the desired distribution [91].

Domain growth and coalescence. We used the previously developed "phantom-nuclei algorithm" (Fig. 2.3) [33]. The algorithm implements the deterministic growth, eliminates passively replicated origins, and joins coalesced domains. In contrast to naive lattice simulations, the algorithm can calculate the state of the genome at any time step without computing intermediate time steps. The routine results in a list of replicated and unreplicated domain lengths that can be used to calculate the different statistics. The most commonly used one in the thesis is the average replication fraction.

Completion time. We use the bisection method and the phantom-nuclei algorithm to search for the first t at which the replication fraction f becomes 1 up to some tolerance level [92]. A faster and more accurate method is to use the phantom nuclei subroutine to eliminate passively replicated origins and calculate directly the coalescence times for each of the utilized origins. The largest coalescence time is then the completion time.

All programming was done using Igor Pro v. 6.22 [93].

Chapter 3

Spatiotemporal Replication Program of Budding Yeast

In this chapter, we develop a parameterized model of the replication program in budding yeast using the theory presented in Chapter 2. While licensing in budding yeast is known to be sequence specific and therefore deterministic in position*, there are opposing views as to whether the initiation time is deterministic or stochastic. To deal with this issue, we use a parameterization that is general enough to describe both deterministic and stochastic initiation. We then perform least-squares fits to a set of recently published time-course microarray data on *Saccharomyces cerevisiae* (budding yeast) [57]. We will see the results show that stochastic effects are important.

From the fits, we extract origin positions, firing-time distributions, origin efficiencies, and a global fork velocity. We find that the wide firing-time distributions suggested by the data are incompatible with a naive deterministic model and are best described by a stochastic process. We also find that the later an origin fires on average, the greater the variation in firing times. In other words, the timing and precision of origin firing are strongly correlated. This correlation and its biological significance is further investigated in Chapter 4. Lastly, we quantify the effects of various experimental imperfections on the fit parameters and argue that these imperfections do not invalidate our major findings. In terms of the three

*Strictly speaking, licensing is never deterministic and always spatially continuous. Thus, the deterministic origin position should be viewed as a simplification that makes the model fitting practical. Compare to other eukaryotes, this simplification is reasonable because the spatially continuous ORC occupancy show well-localized peaks [47].

themes mentioned in Sec. 1.4, this chapter provides an example of the use of mathematical models to extract quantitative information from replication experiments.

3.1 Introduction

The kinetics of DNA replication in eukaryotic cells are carefully controlled, with some parts of the genome replicating early and others replicating later. Patterns of replication timing correlate with gene expression, chromatin structure, and sub-nuclear localization, suggesting that replication timing may play an important role organizing the nucleus and regulating its function [94] (discussed in Chapter 4). The timing of DNA replication is regulated largely by the timing of replication origin initiation. Although the biochemical steps of origin firing are increasingly well-understood (major steps mentioned in Sec. 1.1), the regulation that leads to defined patterns of replication timing is still a mystery [5].

3.1.1 Deterministic or random?

A prevailing picture of eukaryotic replication is that origins are positioned at defined sites and activated at preprogrammed times [95]. In *S. cerevisiae*, origin positions are defined, in part, by 11–17-basepair autonomous replicating sequence (ARS) consensus motifs, which bind the origin recognition complex (ORC) [5]. The timing of origin initiation is more controversial: although microarray experiments have generally been interpreted assuming a deterministic temporal program [20], recent molecular-combing experiments suggest that the initiation of individual origins is stochastic [21, 22]. A more subtle issue is that the consequences of the spatiotemporal connections between origin initiation and fork progression in multiple-origin systems are often not taken into account. An origin site is sometimes implicitly assumed to be replicated solely by the origin itself [39], even though the locus can also be replicated by nearby origins. Our view is that a more rigorous analysis of microarray replication data based on the probabilistic framework presented in Chapter 2 can yield greater insight into the replication process and contribute to forming a more accurate picture of the replication kinetics. We argue that both the population-averaged microarray and the single-molecule combing experiments are compatible with stochastic origin initiation and that the apparent disagreement is resolved after performing a more sophisticated analysis of the microarray data.

3.1.2 Indicator of stochasticity

The microarray experiments yield the fraction of cells in a population that have a specific site replicated after a given time in S phase (see Sec. 3.4 for more experimental details). The origin positions correlate with peaks in the graph of replication fraction vs. genome position, since an origin site is replicated before its neighbouring sites (Fig. 3.1). Following standard nomenclature, we define a median replication time t_{rep} as the time by which half of the cells show replication of the origin locus [20]. Implicit in the traditional interpretation of this type of data are the assumptions that the variation of firing times of an origin, t_{width} (defined as the time for the origin locus to go from 25% to 75% replicated), is small and that t_{rep} is independent of t_{width}.

In more detail, let the replication fraction be $f(x, t)$, where x is the genome position and t is the time elapsed since the start of S phase. For microarray data, $f(x, t)$ represents the fraction of the population that has replicated locus x by time t. Looking at the spatial part of the replication fraction, one expects that an efficient origin that is seldom passively replicated would show an apparent peak at the origin position, as the site is almost always replicated before its surrounding. Thus, peaks in $f(x, t)$ correspond to origin sites. At a fixed x where an origin resides, the peak height then scales with the number of initiations of the particular origin in the population. If an origin always fires at a given time, the corresponding $f(t)$ will resemble a step function. By contrast, if an origin fires stochastically, $f(t)$ will be smooth.

Using 275 previously identified origin sites [96], we extract 275 $f(t)$ curves from the microarray data. (Fig. 3.1B shows one of them.) These $f(t)$ curves are well described by sigmoids with parameters t_{rep} and t_{width}. We used the Hill equation

$$f(t) = \frac{1}{1 + \left(\frac{t_{rep}}{t}\right)^r}, \tag{3.1}$$

where r is related to t_{width} via

$$t_{width} = \left(3^{1/r} - 3^{-1/r}\right) t_{rep}, \tag{3.2}$$

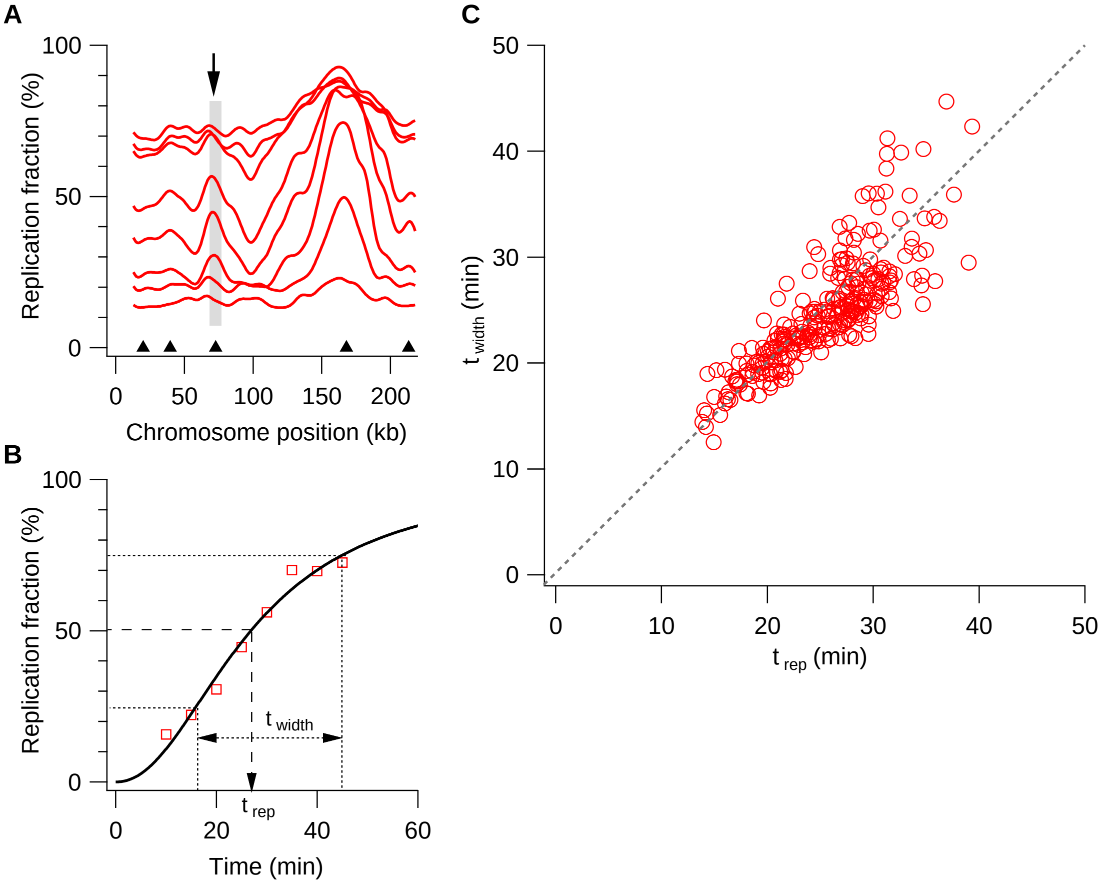


Figure 3.1: A simple analysis of replication timing data. **A**. Smoothed time-course microarray data for Chromosome I. Triangles are origins identified in [96]. The arrow indicates the origin analyzed in B. **B**. Replication fraction of the indicated origin versus time. Fitting the data with the sigmoidal curve defined by Eq. 3.1 gives the median time t_{rep} and the 25–75% time width t_{width}. **C**. Time widths versus median times for all 275 origin loci identified in [96]. The dotted line shows the linear function $t_{width} = t_{rep}$.

to fit the $f(t)$ curves and extracted 275 pairs of t_{rep} and t_{width}. The result is plotted in Fig. 3.1C. Before attempting a more complete analysis, we perform a simple analysis of local replication fraction that ignores the complications due to passive replication.

This simple analysis suggests that, in contrast to the deterministic timing scenario that assumes t_{width} is much less than t_{rep}, the extracted t_{width} approximately equals t_{rep} (Fig. 3.1C). The apparent variability in origin timings suggests that stochasticity is important for an accurate description of the replication program. Moreover, in a model where the timing of origin firing is regulated by external triggers [94] (see Sec. 4.1), one expects no correlation between t_{rep} and t_{width}. In other words, the t_{width} points would scatter about a horizontal line in Fig. 3.1C, implying that variations in t_{width} are independent of t_{rep}. Instead, we see a strong correlation between t_{width} and t_{rep}, suggesting the two are mechanistically related.

3.2 Fitting time-course replication fraction with the sigmoid model (SM)

The discrepancies between a naive, deterministic picture of origin initiation and the data motivate a more detailed approach. Based on the framework developed in Chapter 2, we introduce an analytical model, the "sigmoid model" (SM), that can generate replication fraction $f(x,t)$ to fit the genome-wide time-course microarray data in [57].

The framework in Chapter 2 links $f(x,t)$ with the spatiotemporal initiation rate $I(x,t)$ via Eq. 2.5. Since the origins in *S. cerevisiae* are well-localized, the genome-wide $I(x,t)$ can be separated into contributions from each origin. Each origin is then described by its own firing-time distribution $\phi_i(t)$, defined in Eq. 2.11 as the probability that origin i at site x_i initiates between time t and $t + dt$. The use of firing-time distribution instead of an initiation rate $I_i(t)$ simplifies the calculation of origin efficiency via Eq. 2.13. The two quantities are interchangeable via $I_i(t) = \phi_i(t)/[1 - \Phi_i(t)]$ (Eq. 2.12). Inspired by the sigmoidal replication fraction profile in Fig. 3.1B, we assign each origin a sigmoidal cumulative firing-time distribution $\Phi_i(t)$—hence the name "sigmoid" model.

Below, we detail the parameters used in the SM fits. The model parameters can be separated into a "local" group that describes the properties of each origin and a "genome-wide" group that describes quantities that are roughly constant across the genome. Within the local group, there are two types of parameters, one for the position and the other for the firing-time distribution. These parameters are local in space (genome position) but global in time (throughout S phase):

Origin position. Each origin is associated with a unique location x_i along the genome. Initial guesses of the x_i include all 732 origins recorded in the OriDB database [97]. We count origins that were less than 5 kb apart as a single origin throughout the fitting process, as the data (resolution ≈ 2 kb) cannot be used to distinguish between a single efficient origin and a group of less-efficient origins (see Sec. 3.4 and the discussion around Eq. 2.14). Before attempting the genome-wide fit, we first fit each chromosome separately to eliminate false origins and origins that do not contribute enough to the replication fraction. This allows the genome-wide fit to run more smoothly. The criteria for elimination are:

1. Mode of firing-time distribution $<$ 10 minutes (min) and efficiency defined by Eq. 2.13 $<$ 0.5.

2. Cumulative firing-time probability $<$ 0.3 at end of S phase (estimated to be 60 min).

The first criterion aims to identify false-origin peaks in microarray data that originate from contamination instead of origin activity. (Small fragments of unreplicated DNA along with A-T rich sequence can be mistaken as replicated DNA, as mentioned in Sec. 3.4.) Contamination produces microarray peaks in the early but not the mid and late time points. In the model, this is effectively the same as an origin that fires very early (before the first time point at 10 min) but is inefficient. The second criterion simply eliminates origins that do not contribute much throughout S phase.

Firing-time-distribution parameters. The cumulative firing-time distribution $\Phi(t)$ has two parameters: the median time $t_{1/2}$ at which $\Phi(t = t_{1/2}) = 0.5$ and the width t_w, defined as the difference $t_{0.75} - t_{0.25}$. Explicitly, $\Phi(t)$ takes the form of Eq. 3.1, but with t_{rep} and t_{width} replaced by $t_{1/2}$ and t_w, respectively (see Appendix 4.A.1 for a discussion on the use of Hill equation). This function is a valid cumulative initiation probability function, satisfying the constraints $0 \leq \Phi(t) \leq 1$, $\Phi(t < 0) = 0$, and $d\Phi(t)/dt \geq 0$ given after Eq. 2.5. We note that these two parameters $t_{1/2}$ and t_w differ from t_{rep} and t_{width} in Sec. 3.1.2. The former pair describes the “firing time” of an origin, while the latter pair describes the “replication time” of an origin site, which is replicated by both the on-site origin and nearby origins.

There are three types of genome-wide parameters:

Fork velocity v. We use a velocity v that is constant across the genome and throughout S phase. There are three justifications to the use of such a simple velocity: First, a constant velocity fits the genome-wide data well. Second, fitting the data with a model that uses spatially varying fork velocities does not improve the fit much (Appendix 3.A.2). Third, there is experimental evidence that the fork velocity is remarkably uniform across the genome [48].

Background b_g. The reason for introducing this parameter comes from the observation that the 10-minute-time-point data do not start close to f=0 (see Sec. 3.4). While introducing a variable background is possible, we find that a simple constant background is sufficient for the global fit. This parameter is global both in space and time.

Normalization factors α_t. These parameters correct for various artifacts that cause the microarray data to not cover the entire range of possible fractions, 0% to 100% (see Sec. 3.4). We propose a genome-wide normalization factor α_t for each time point. Combining the normalization factors with the background parameter, we generate a modified replication fraction

$$f_{mod}(x,t) = \alpha_t \left[f(x,t) + b_g \right] \tag{3.3}$$

to fit the data. Notice that, in Eq. 3.3, α_t is global in space but not time, while b_g is global both in space and time.

In summary, the SM parameters include origin position, the $t_{1/2}$ and t_w that describe firing-time distributions, fork velocity, and factors to deal with experimental artifacts (discussed in Sec: 3.4). Using this model, we performed least-squares fits to the time-course microarray data in [57]. The fit was done using the Global Fit package in Igor Pro 6.1 (Wavemetrics Inc. `http://www.wavemetrics.com`). We fit to the unsmoothed microarray data (see Appendix 3.A.3; Fig. 3.16). To speed up the code, we wrote an external C-language code that the fit program calls to do key function evaluations. On a personal computer with an Intel Core2 Duo CPU @ 3.16 GHz, the fit for an average chromosome ($\approx$ 900 points and $\approx$ 150 parameters) takes about 5 minutes. The time to fit the entire genome of *S. cerevisiae* is about 10 hours[*]. All local parameters of the SM are tabulated in Supplementary Table I; global ones are in Supplementary Table III (see Appendix 4.A.2). The fit to Chromosome XI is shown in Fig. 3.2 to demonstrate that the model captures the replication process well. Fits to all chromosomes can be found in Fig. 4.8 at the end of Chapter 4. A detailed statistical analysis of the fits is presented later in Appendix 3.A.1.

[*] A rough empirical estimate of the time to do fits with different number of data points N_d suggests that the fit time scales with N_d^2.

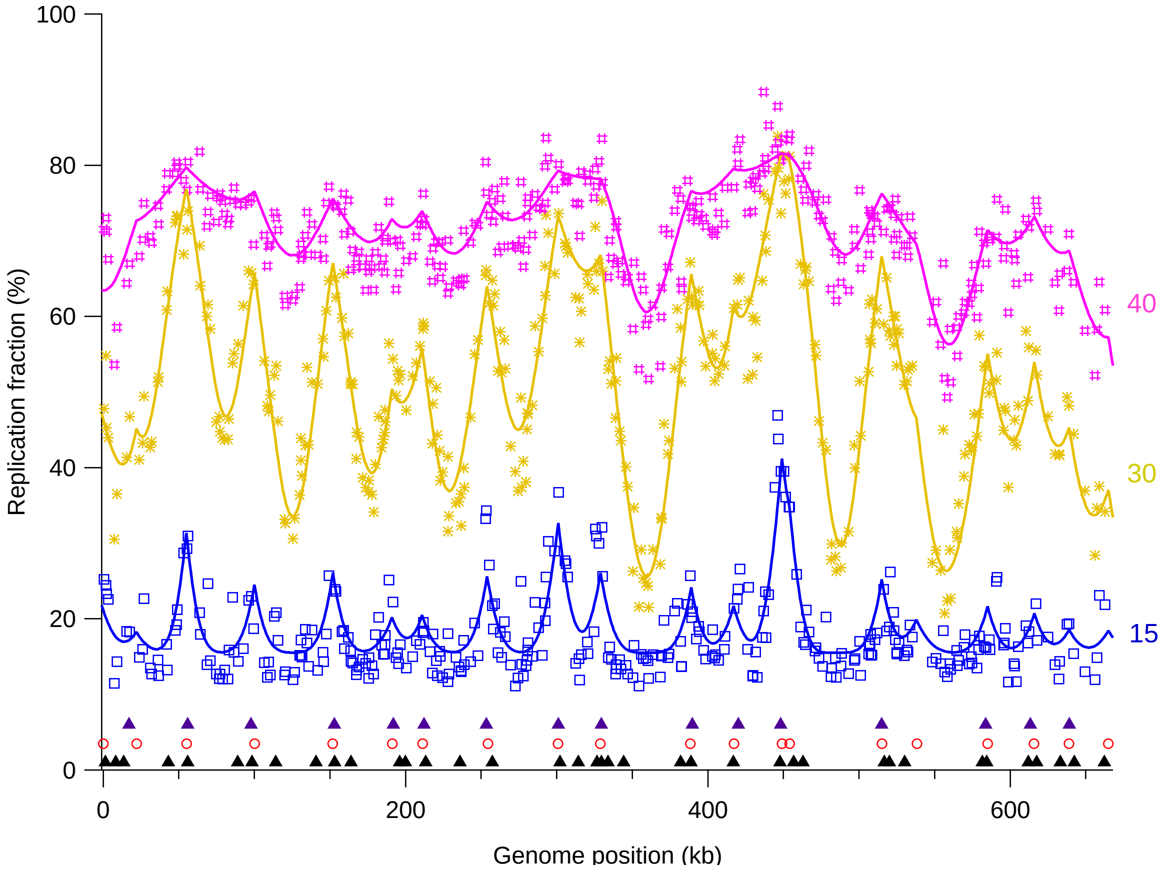


Figure 3.2: Chromosome-XI section of the genome-wide fits. Markers are data; solid lines are fits from SM. At the bottom, upper row of solid triangles denote origin positions identified in [96]; middle row of open circles denote the estimated origin positions in the sigmoid model (SM); and the lower row of triangles correspond to origins in the OriDB database [97]. The three curves from bottom to top correspond to the replication fraction $f(x)$ at 15, 30, and 40 minutes after release from the *cdc7-1* block. The dataset covers the genome at ≈2-kb resolution.

3.3 Results

3.3.1 Extracted origins and fork velocity

The list of initial guesses for origin positions consisted of 732 positions from the OriDB database that had been previously identified using a variety of methods [97]. All results presented below do not depend sensitively on this initial list, since we allowed origin positions to vary in the fit. After eliminating origins according to the criteria described above, the SM gave 342 origins (origin parameters tabulated in Supplementary Table I; see Appendix 4.A.2). Of the 342 origins, 236 colocalize with the 275 origins identified by Alvino *et al.* using a similar dataset [96]. The remaining 106 origins were not reported in [96] because they do not associate with apparent peaks in the microarray data. We find that 75% of the 106 colocalize (within 5 kb) with origins in the OriDB database [97].

The genome-wide fork velocity we extracted from the SM is 1.9 kb/min. Our statistical analysis of the fits suggests an error of 0.2 kb/min on v (see Appendix 3.A.1). Consistent with this conclusion, recent work that monitored the movement of GINS, an integral member of replication forks, showed that the fork progression rate is 1.6 ± 0.3 kb/min and does not vary significantly across the genome [48]. Our conclusion and that in [48] contrast with a previous analysis, where variations in slope from peak to trough in the replication fraction were interpreted as fork-velocity variations [20]. In our model, these variations are mostly accounted for by variations in firing-time distribution and the levels of passive replication [42].

3.3.2 Firing-time distributions and initiation rates

The least-understood aspect of the eukaryotic replication process is its temporal program [5]. One reason is that there has been no direct way of visualizing both the spatial and temporal aspects of replication at high resolution. Another reason is that the implications of temporal stochasticity in initiations have often been neglected. We extracted the firing-time distributions of the 342 origins identified in the SM (Fig. 3.3A). The widths are comparable to the length of S phase, confirming that stochastic effects play an important role. The kinetic curves of the origins extracted from the model also show that t_w increases with $t_{1/2}$ (Figs. 3.3 and 3.4). Further, we show that four representative, previously studied origins

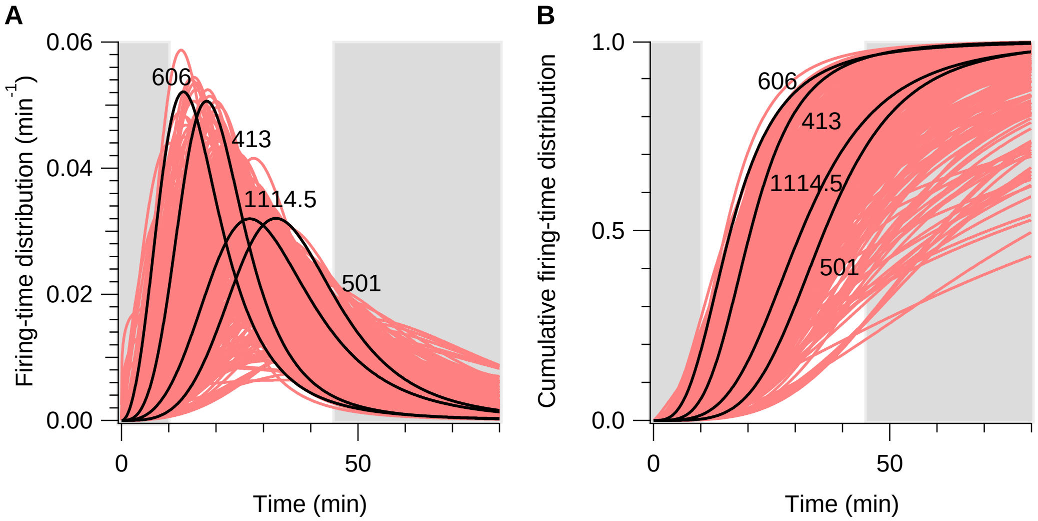


Figure 3.3: Firing-time distributions. **A**. Firing-time distribution for the 342 origins extracted using the SM. Darker curves are representative distributions chosen to illustrate origins with different values of $t_{1/2}$. Numbers denote their ARS number in the OriDB database [97]. Grey background corresponds to times not included in the data (t <10 min and t >45 min); curves in these domains are extrapolated. The length of S phase is roughly 60 min*. **B**. Cumulative firing-time distributions $\Phi(t)$ for the 342 origins. Each trace is the kinetic curve that would be measured for the corresponding origin site if there were no neighbouring origins. Darker curves correspond to darker curves in A. The potential efficiencies of the origins are taken to be the value of $\Phi(t)$ at $t = 60$ min.

(ARS 413, ARS 501, ARS 606, and ARS 1114.5) are typical origins that follow this global $t_{1/2}$-t_w relationship (Fig. 3.4).

The rate of initiation, defined as the number of initiations per time per unreplicated length, is a crucial parameter in kinetics (see Chapter 2). It has been proposed that an increasing rate of initiation later in S phase would lead to robust completion of replication, even when origin firing is stochastic [99]. To investigate the initiation rate in the SM, we plotted initiation rates averaged over the genome and over individual chromosomes (Fig. 3.5). Our results show the initiation rate rising for most of S phase and then declining in late S phase. A similar pattern has been described in a number of organisms [38]. However, the genome-wide-averaged initiation rate we extracted does not decay to zero before S phase ends, as it does in the analysis in [38]. In [38], the use of t_{rep} rather than the full distribution of firing times of an origin leads to an underestimation of origin initiation at

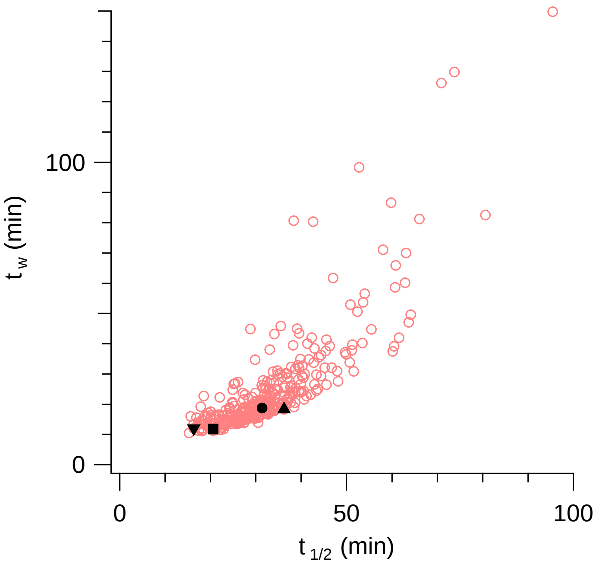


Figure 3.4: Trend of $t_{1/2}$ vs. t_w. Markers are results from the SM. Solid square, triangle, inverted triangle, and circle correspond to ARS 413, 501, 606, and 1114.5, respectively.

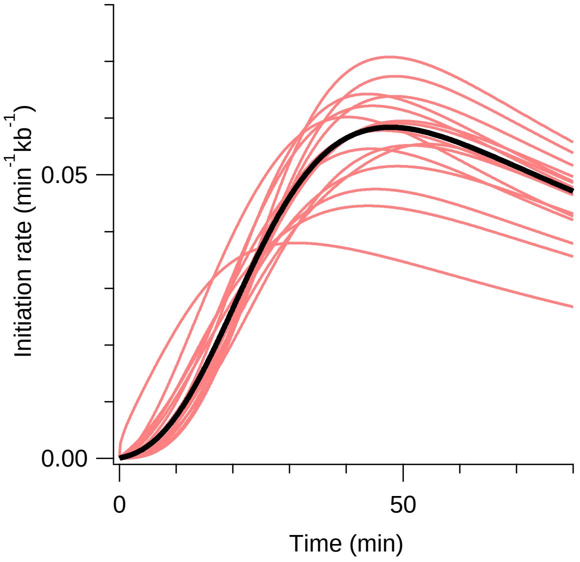


Figure 3.5: The heavy dark curve is the genome-wide averaged initiation rate, while the lighter curves are chromosome-averaged rates.

late times. Nonetheless, since the proposed universality of the initiation rate across species was for *scaled* initiation rates, that conclusion may well survive reanalysis of all datasets with a stochastic model, which would modify the extracted average rates and scalings for all the analyzed organisms.

3.3.3 Origin efficiencies and passive replication

Understanding origin efficiency is important because these efficiencies determine the replication-completion time and the robustness of the replication program [40, 41]. The efficiency of an origin is closely related to its geometry—its location relative to other origins. Imagine two highly efficient origins placed near each other; only one of the two origins will fire in any given cell because the initiation of one origin will passively replicate the other origin. Placing an efficient origin next to an inefficient one should decrease the firing of each by different amounts. For an isolated origin, one expects that its efficiency would be unaffected by passive replication.

In our analysis, we distinguish between "efficiency," which is traditionally defined as the probability that an origin fires during normal replication, taking into account passive replication effects, and "potential efficiency." The latter term is defined as the probability that an origin would have initiated before the end of S phase if there were no passive replication [100]. We will occasionally write "utilized efficiency" instead of just "efficiency" to emphasize the effect of passive replication. Experimentally, techniques that hinder fork progression to avoid passive replication have been applied to extract the potential efficiencies of origins [56]. However, such approaches provide only rough estimates of potential efficiency because the fork-stalling drug also blocks origin firing in late S phase. Thus, to determine whether the utilized efficiency of origins is due primarily to their potential efficiency or to their proximity to neighbouring origins, we investigated the relationship between efficiency and potential efficiency.

We use the extracted origin positions and firing-time distributions to calculate both the utilized efficiencies (via Eq. 2.13) and potential efficiencies of all identified origins (via evaluating $\Phi_i(t = t_{\text{end}})$; see Fig. 3.3B caption and Sec. 2.1.2). Here, t_{end} is the length of S phase, and we estimate it to be 60 minutes from flow-cytometric determination of replication progress (see Sec. 3.4). We found that more than half of the origins have high

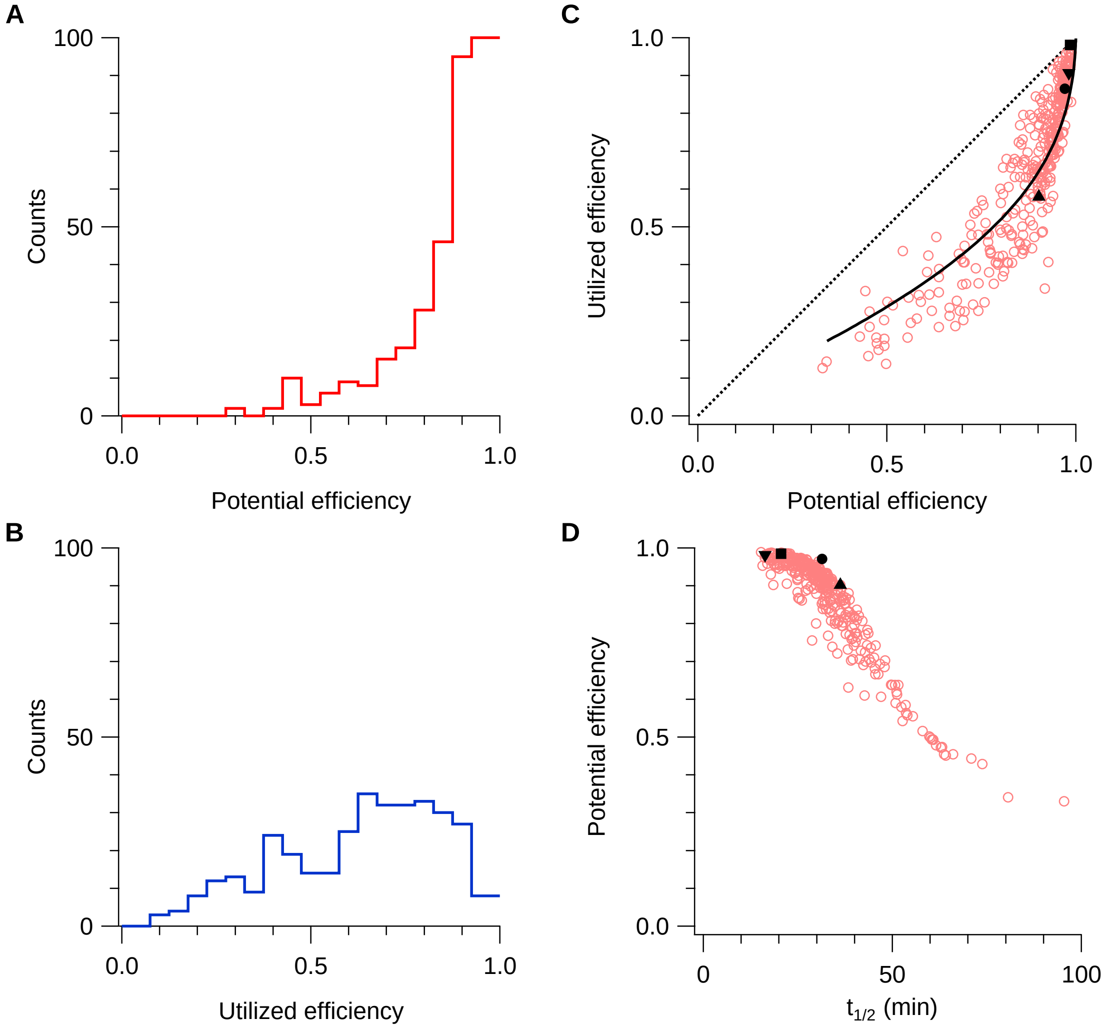


Figure 3.6: Origin efficiencies. **A**. Histogram of potential efficiencies with bin width = 0.05. More than half of the origins have high potential efficiency (> 0.9). Median value = 0.91. **B**. Histogram of utilized efficiencies with bin width = 0.05. Median value = 0.68. **C**. Utilized efficiency versus potential efficiency. Markers would lie on the dotted line if there were no passive replication. Solid markers correspond to the same origins in Fig. 3.4. The solid curve is a mean-field calculation (see Sec. 3.3.4). **D**. Potential efficiency vs. $t_{1/2}$. There is roughly a one-to-one relationship between the potential efficiency and $t_{1/2}$.

potential efficiency (> 0.9), but the utilized efficiencies of origins vary much more (compare Fig. 3.6B to A). Furthermore, we found that the relation between utilized efficiency and potential efficiency can be approximately accounted for by a mean-field argument (see Sec. 3.3.4), where all neighbouring origins are replaced by an average neighbour that has the genome-wide-averaged firing-time distribution (Fig. 3.6C). Thus, the efficiency of origins in *S. cerevisiae* can largely be explained by geometric effects due to neighbouring origins.

3.3.4 Mean-field analysis of origin efficiency

The relationship between efficiency and potential efficiency shown in Fig. 3.6 can be mostly explained by a mean-field analysis. The idea is that all the neighbouring origins of an origin are replaced by an "average neighbour" whose firing-time distribution is the average of all the distributions. We averaged over all 342 firing-time distributions in the SM to produce the genome-wide-averaged $\phi_{avg}(t)$. We then computed the average nearest-neighbour distance ($\approx$ 28 kb) to locate the average neighbour. Next, we approximated t_w as a function of $t_{1/2}$ by fitting a power-law through Fig. 3.4 ($t_w \approx 11 + 0.002\, t_{1/2}^{2.4}$). The analytic relationship between t_w and $t_{1/2}$ implies that the potential efficiency is also a smooth function of $t_{1/2}$. Finally, the efficiency was calculated by placing the average neighbour at the average nearest-neighbour distance beside origins. Going through all the $t_{1/2}$ values extracted, we generated the curve shown in Fig. 3.6C. This analysis suggests that the geometric effect we see on utilized origin efficiency is not specific to the particular arrangement of origins in budding yeast and would be generally expected for a genome with this density of origins.

3.3.5 Case study of origin ARS501

As seen in Fig. 3.6D, later-firing origins have lower potential efficiency. The monotonic decrease in efficiency with increasing $t_{1/2}$ is a consequence of the t_w-vs.-$t_{1/2}$ relationship mentioned above. The larger t_w values associated with later-firing origins imply that their chance of initiating before S phase ends is less than that of earlier-firing origins.

A previous experiment reported that although the ARS501 origin is late firing, its kinetics (replication fraction $f(t)$ curve) and efficiency resemble that of an early-firing origin [101]. The ARS501 kinetics was used to support a scenario where origin initiation is regu-

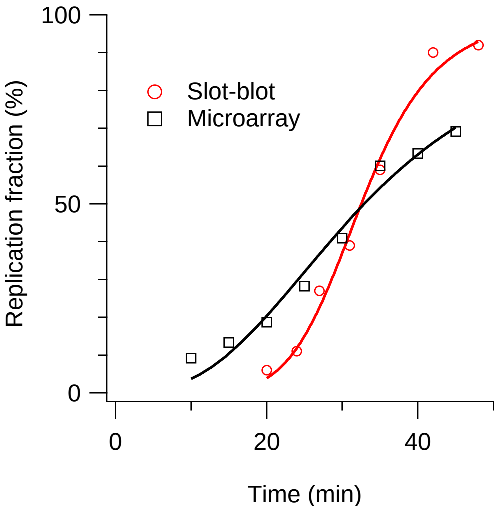


Figure 3.7: Replication fraction of ARS501. ARS501 is located on Chromosome V at ≈ 549 kb. Circles are data from a slot-blot experiment [101]; squares are data from the newer microarray experiment [57]. Lines are fits to the data using a sigmoid (Hill equation in Eq. 3.1). Values for t_{rep} and t_{width} are extracted for comparison. For the slot-blot, $t_{rep} = 33$ min and $t_{width} = 11$ min. For the microarray, $t_{rep} = 33$ min and $t_{width} = 26$ min.

lated deterministically in time. We compared the kinetic curves of ARS501 obtained from the earlier slot-blot experiment [101] to that from the present microarray data [57] and found that the two curves differ significantly[*]. The microarray kinetic curve suggests that ARS501 is a typical origin with $t_{rep} \approx 33$ min and $t_{width} \approx 26$ min, while the slot-blot data suggest $t_{rep} \approx 33$ min and $t_{width} \approx 11$ min (Fig. 3.7). For comparing origin properties, we argue that the microarray data are more reliable than the slot-blot data because the former has information about the relative behaviour of origins genome wide, while the latter contains curves for only a few sites. Simply stated, microarray data contain more relative information about sites than the slot-blot data. According to our analysis, ARS501 is neither unusually efficient (utilized efficiency ≈ 0.58) nor unusually late ($t_{1/2} \approx 36$ min; the median and standard deviation of the 342 $t_{1/2}$ values are 31 min and 11.2 min, respectively).

[*] The slot-plot experiment in [101] is essentially the same as the microarray experiment except that it can probe only one genome position.

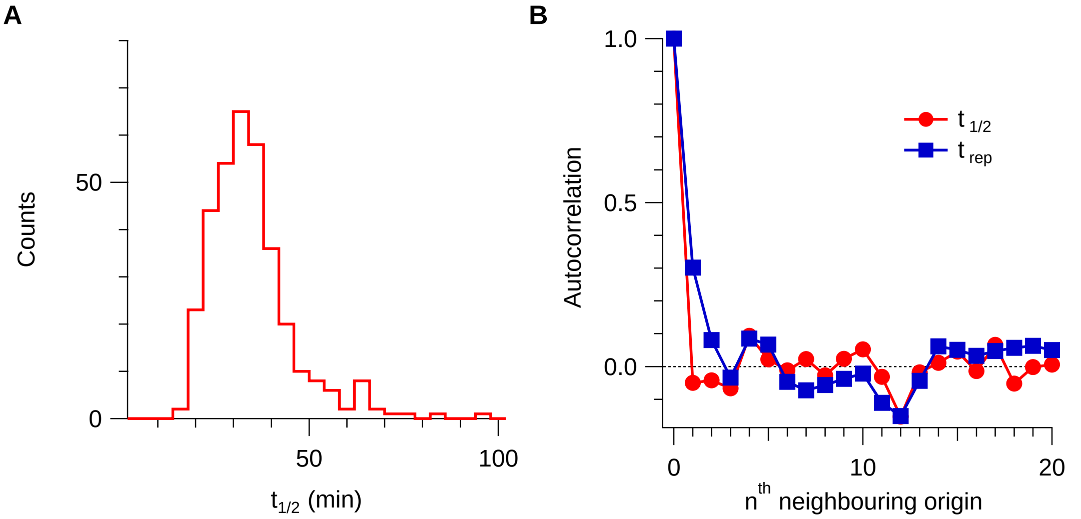


Figure 3.8: Histogram and autocorrelation of origin firing time. **A**. Histogram of $t_{1/2}$ with bin width = 4 min. There is no sharp distinction between early and late origins. Median value = 31.0 min, standard deviation = 11.2 min. **B**. Autocorrelation of $t_{1/2}$ and t_{rep}. The graph shows the $t_{1/2}$ and t_{rep} correlation between an origin and its n^{th} neighbour. The range of the x-axis corresponds to roughly 700 kb of genome distance, as the average interorigin distance extracted is 35 kb.

3.3.6 Initiation clustering?

The microarray replication profiles $f(x,t)$ show that different parts of the genome replicate at different times. McCune *et al.* studied a mutant yeast strain where origin firings are largely limited to the first half of the cell cycle and found that the typical replication fraction of some regions is unaltered while it decreased in other regions [57]. They thus hypothesized that there are relatively long stretches of chromosomes that replicate early and late and that temporally alike origins are clustered together to form these early- or late-replicating regions [57].

To investigate this hypothesis, we note that the distribution of $t_{1/2}$ has a single peak (Fig. 3.8A; also seen in [20]), suggesting a continuum of firing times and arguing against distinct categories of temporally alike origins (e.g., early and late origins). To test the spatial aspect, we calculated the autocorrelation function of the origins' $t_{1/2}$ values (Fig. 3.8B) [92]. In this test, a positive value means that, typically, an origin and its n^{th} neighbour are likely to have $t_{1/2}$ that are both larger or both smaller than the average firing time, as would be expected if temporally alike origins were to cluster. Conversely, negative values would mean that an origin and its n^{th} neighbour were likely to have $t_{1/2}$ values that were anti-correlated. We found that the autocorrelation function fluctuates around zero for $t_{1/2}$ but is clearly non-zero at the nearest neighbour for t_{rep} (Fig. 3.8B). The former result indicates that the intrinsic initiation time among *origins* is independent. The latter result indicates that the time at which neighbouring *origin sites* are replicated is correlated. The correlation arises because a fork from one origin sometimes passively replicates the site of a neighbouring origin. Thus, the observation that neighbouring origin sites tend to replicate at similar times is consistent with the inference that neighbouring origins initiate independently.

3.3.7 Robustness of the replication program

The difference between the distributions of efficiency and potential efficiency gives insight about the robustness of the replication program. Although the utilized efficiencies varies over a wide range, most of the origins have high potential efficiencies, implying that there are more potential origins than needed to replicate the genome and that many potentially efficient origins appear to be dormant. In circumstances where some forks stall because of DNA breakage or other replication stress, these normally dormant yet highly efficient origins, not being passively replicated, would initiate to help replicate the DNA [41, 102]. In this way, redundancy is built into these potentially efficiency origins to safeguard the replication process against replication stress.

Furthermore, Fig. 3.5 shows that the initiation rate extracted using the SM increases throughout most of S phase. Biologically, this design allows any large, unreplicated regions that remain toward the end of S phase to be replicated with increasing probability, ensuring timely completion of genome duplication [73]. An increase in the origin-initiation rate as S phase progresses has been observed in every dataset that has been studied, and several plausible mechanisms for such an increase have been suggested [37, 38, 100].

3.3.8 Summary of results

Among the various results presented above, the most important one is the genome-wide correlation between $t_{1/2}$ and t_w (Fig. 3.4). As mentioned above in Sec. 3.3.5, this correlation implies that the earlier-firing origins are also potentially more efficient. Then, in Sec. 3.3.4, we established that the utilized efficiency can be approximated as a function of potential efficiency and the density of origins. This suggests that the utilized efficiency is also partly a consequence of the $t_{1/2}$-t_w correlation. Since the distributions of efficiency and potential efficiency are largely determined by the correlation, the robustness against replication stress and for timely duplication discussed in Sec. 3.3.7 is yet another consequence of the $t_{1/2}$-t_w correlation. This important feature of the replication program is the focus of Chapter 4.

3.4 Discussion of data limitations

Although the microarray experiments analyzed here provide high-quality data, artifacts and limitations should be addressed. We devote this section to show that our major findings remain valid in the presence of these issues. We begin with a brief description of the experiment, following [96] and [20]. The data we analyzed [57] were obtained using similar procedures.

In a protocol inspired by the classic Meselson-Stahl experiment [103], budding yeast cells were grown in an isotopically dense (^{13}C, ^{15}N) medium for a few generations at 23 $^\circ$C and then synchronized at G1 by exposure to alpha mating pheromone. The culture was then resuspended in an isotopically light (^{12}C, ^{14}N) medium and further synchronized at the G1/S boundary by incubation at 37 $^\circ$C, the restrictive temperature for *cdc7-1*. When 93% of the cells were budded (a sign that the cells are not dead), the temperature was lowered to the permissive temperature 23 $^\circ$C to allow cells to enter S phase. Newly replicated DNA would incorporate the light isotope to form a heavy-light double-strand, while unreplicated DNA would be heavy-heavy. Samples were collected throughout S phase. The DNA of the collected cells was first fragmented with a restriction enzyme (Eco RI). Heavy-heavy and heavy-light DNA were then separated by ultracentrifugation, separately labelled with Cy3-dUTP and Cy5-dUTP, and hybridized to an open-reading-frame microarray*. The intensities, after normalization by the mass of the sample, were used to calculate the fraction of replication [96].

*An open reading frame (ORF) is a DNA sequence that does not contain a stop codon (possible stop codons are TAA, TAG, and TGA). An ORF microarray uses these sequences as spatial probes. In budding yeast, there is on average 1 ORF per 2 kb, hence the spatial resolution of $\approx$ 2kb.

3.4.1 Asynchrony in cell population

As shown in Fig. 1.3, asynchrony in a cell population in general denotes that each cell is at a different stage of the cell cycle. In the discussion below, we use asynchrony to specifically describe the fact that cells in a population enter and progress through S phase at different times and with different rates. It is apparent that asynchrony widens firing-time distributions. Consider a scenario where the timing of every origin is deterministic. Since cells in an asynchronous culture enter S phase at different times, the initiation times will appear to be stochastic. To assess the effect of asynchrony on the parameters we extracted, we extended our formalism to include asynchrony.

For the modelling, we first distinguish between "starting-time asynchrony" and "progressive asynchrony." For the microarray experiment analyzed, the cell culture was synchronized in two steps (first by alpha-factor incubation, then with *cdc7-1* block) before samples were taken for hybridization. We define starting-time asynchrony in terms of the distribution of release times relative to the last synchronization procedure. In other words, this is the asynchrony that reflects imperfections in the procedure used to synchronize the cell cycles of the population under study. Now, consider a scenario where the synchronization procedures produce a perfectly synchronized cell culture. If the replication program is not strictly deterministic, the DNA content for each cell will evolve differently as S phase proceeds. This "progressive asynchrony" is a consequence of the stochastic replication program. The probabilistic model presented in Chapter 2 already captures the effects of progressive asynchrony on the replication fraction. Since the data analyzed contain both types of asynchrony, we extend the formalism to include starting-time asynchrony.

We model the starting-time asynchrony of a cell population by a starting-time distribution $\psi_a(\tau)$, defined as the fraction of cells in the population that enters S phase between lab time τ and $\tau + \delta\tau$. For a $\psi_a(\tau)$ centred around $\tau = 0$, cells associated with negative τ enter S phase τ minutes after the nominal start of S phase at $t = 0$. If the probed cell culture has a starting-time distribution $\psi_a(\tau)$, the measured replication fraction profile can be approximated as the convolution

$$f_a(x,t) = \int_{-\infty}^{\infty} f(x,\tau)\psi_a(t-\tau)d\tau, \tag{3.4}$$

with $f(x, t)$ being the replication fraction profile for a perfectly synchronous cell culture $[\psi_a(\tau) = \delta(\tau)]^*$.

We simulated the replication fraction profiles that contain both types of asynchrony using the simulation method described in Chapter 2. The theoretical approximation matches the simulation data well (Fig. 3.9A). The most apparent effects of the starting-time asynchrony are the "pulling" of the peaks in the replication-fraction axis and the "stretching" of the peaks in the position axis. The former mainly translates into an apparently wider firing-time distribution, whereas the latter translates into an apparently faster fork progression rate.

To apply Eq. 3.4 to our analysis, we need to estimate the starting-time distribution. To our knowledge, although there are works that estimate the starting-time distribution resulting from alpha-factor synchronization [104, 105], there are none related to the *cdc7-1* block. Since the *cdc7-1* block is the final synchronization step taken and since it blocks cells at the G1/S boundary, it is important for the estimate of $\psi_a(t)$ to include the effects of *cdc7-1*. To do this, we analyzed the flow-cytometric determination of DNA content (Fig. 1A in [96]).

We measured the width by measuring the spread of DNA content at half the peak height (Fig. 3.10). The width at 0-min is a reference width corresponding to perfect synchrony, as all the cells have 1C amount of DNA. The width at 20-min includes both types of asynchrony and can be used to generate an upper bound of the starting-time asynchrony. A simple image analysis shows that the width of the 20-min peak is 5 pixels larger than that of the 0-min peak. DNA content increases from 1C to 2C over 75 pixels. Using a crude estimate that DNA content linearly increases with the progression of S phase, we converted 5 pixels to 4 min (via 60 min/75 pixel). Since the flow-cytometric peaks are approximately Gaussian, we set $\psi_a(t)$ to a normal distribution with mean zero and standard deviation 2 min, denoted by N(0,2). The estimated asynchrony implies that 95% of the cells would have entered S phase within 8 min of the start of S phase.

We refit the SM to Chromosome-XI part of the data with Eq. 3.4 (instead of Eq. 2.5) and the estimated asynchrony. We found that the local parameters extracted with asynchrony

[*] To fully model the asynchrony, one has to model how the replication-fraction distribution of the cell culture evolves as replication proceeds. This is analytically difficult because it involves calculating all the moments of the replication fraction, of which $f(x, t)$ is only the first. For the analysis here, the approximation in Eq. 3.4 suffices, and the full calculation is not needed.

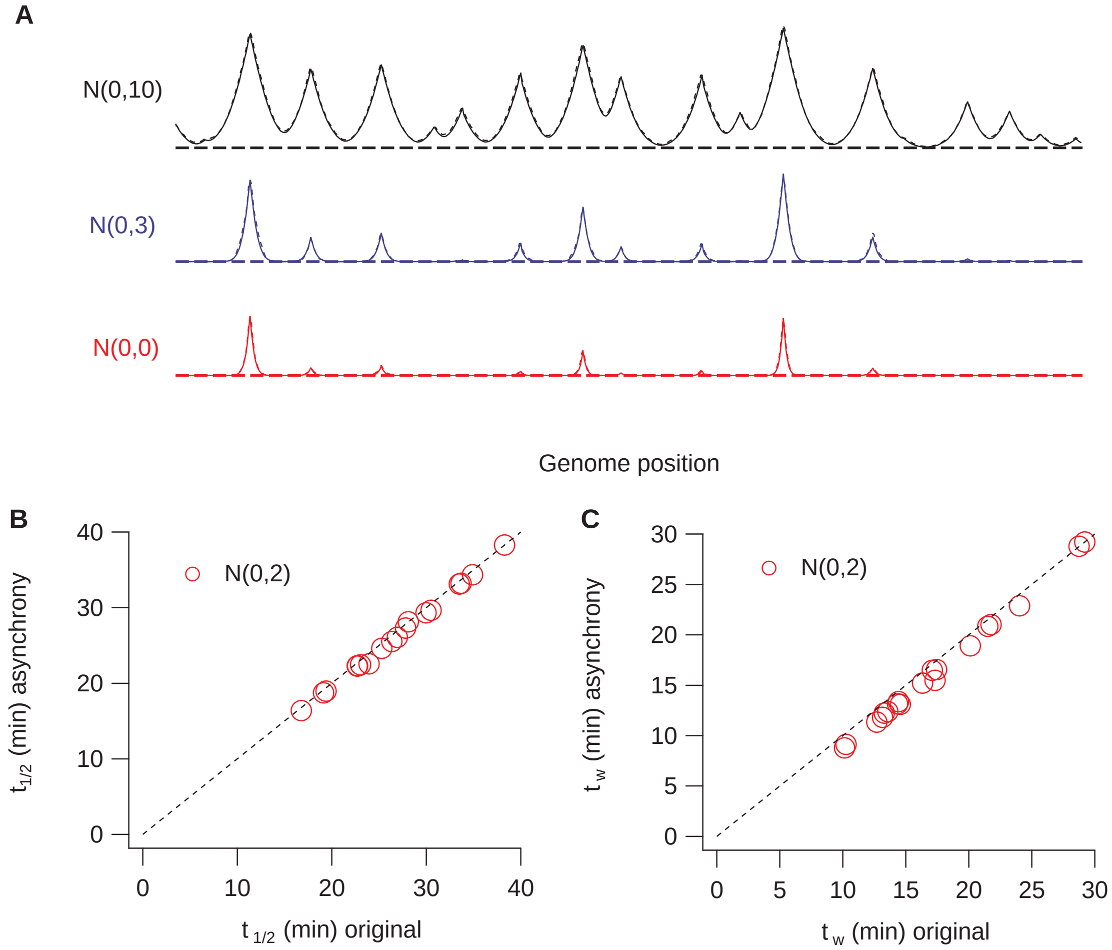


Figure 3.9: Effects of starting-time asynchrony on replication fraction. **A**. Simulation and theoretical replication fraction profile $f(x,t)$ with three different starting-time distributions. The notation N(μ,σ) denotes a normal distribution with mean μ and standard deviation σ. The three curves are generated using the same set of SM parameters (x_i, $t_{1/2}$, t_w and v) and correspond to the same time point. The only difference among them is the starting-time distribution. The theoretical calculation (solid curves) matches the simulations (dashed curves) well. Horizontal dashed lines are the replication fraction 0-lines for the three cases. **B**. Comparison of $t_{1/2}$ fit parameters. The x-axis corresponds to the SM parameters extracted without consideration of asynchrony; the y-axis corresponds to the case with consideration of asynchrony. Dashed line shows $y = x$. **C**. Comparison of t_w fit parameters. The x-axis, y-axis, and dotted lines are as described in B.

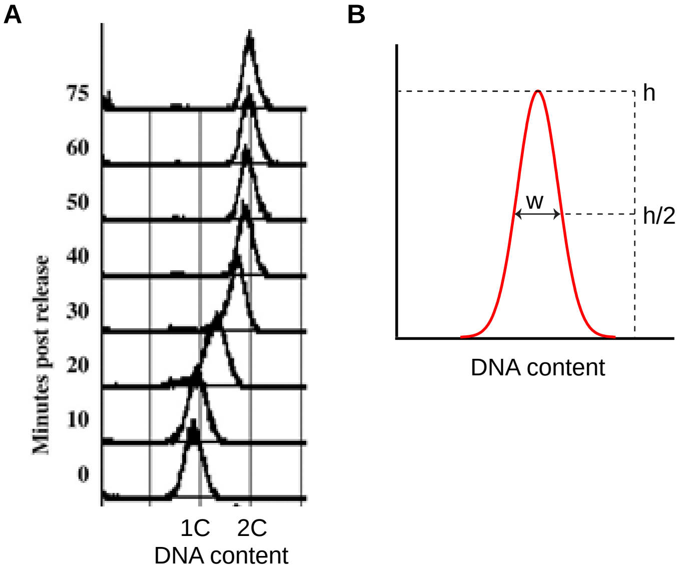


Figure 3.10: Determining the width of DNA content histograms. **A**. The DNA content histograms obtained by flow cytometry after releasing the culture from the *cdc7-1* block. The plot is reproduced with permission from Mol. Cell. Biol. [Vol. 27, pages 6396–6404, Fig. 1A] [96] (Copyright ©2007, American Society for Microbiology). **B**. The width w is defined as the width of the histogram at half the peak height $h/2$.

are not significantly different from those extracted without asynchrony (Fig. 3.9B and C). The estimated starting-time asynchrony shifts $t_{1/2}$ by ≈ -0.5 min, t_w by ≈ -1 min, and v by ≈ -0.3 kb/min. These shifts do not change the relationship between $t_{1/2}$ and t_w, and the results presented in the main text remain valid. We note that using a linear relationship between DNA content and time underestimates the asynchrony; however, refitting using N(0,4) also gives shifts that are unimportant.

3.4.2 Data resolution, data range, data statistics, and S-phase duration

Data resolution. A limitation of the data is its resolution. The dataset covers roughly the entire genome at time points from 10 to 45 minutes, as measured from the release of the *cdc7-1* block. It comprises 8 time points (with 5-minute temporal resolution) and, on average, 6149 position points for each time point (spatial resolution = genome size /

number of points $\approx$ 12000 kb / 6149 $\approx$ 2 kb). The average spatial resolution of 2 kb cannot resolve every single origin. In our fits, the elimination of origins that are less than 5 kb apart from their neighbours reflects this limitation. Given that the fork speed $v \approx 2$ kb/min, a typical origin can cover roughly 60 kb (average $t_{1/2} \times v$) of DNA. Thus, treating all origins in the region $x_i \pm 2.5$ kb as an effective origin at x_i would not change the replication fraction profiles. The average error on the x_i is 0.7 kb for the SM.

The exact number of effective origins that we found depends on the elimination criteria (see Sec. 3.2), as some origins made only marginal contributions to the replication fraction profiles. The parameters of these origins have relatively large errors (t_w± 50% for the SM). They were also sensitive to the form of data used in the fit (e.g., smoothed data vs. raw; Appendix 3.A.3). The marginal origins constitute about 10% of all origins identified. Since they do not affect the replication kinetics significantly (the replication fraction profile at 30 min changes by less than 5%), uncertainties about their numbers and parameters do not change the results presented above.

Data range. Another issue is that the dataset does not cover the entire range of possible replication fraction (0–100%); roughly all the data points spread between 10–90% (Fig. 4.8). One contribution to this artifact is the inability to cleanly separate the replicated fragments from the unreplicated. Alvino *et al.* reported that small fragments and A-T rich sequences of unreplicated DNA are less dense and are physically similar to the replicated fragments [96]. This leads to non-zero replication signals everywhere, even when no DNA is replicated. To understand the upper bound of 90%, we note Alvino *et al.* reported that, for each time point, they normalized the microarray signals by the ratio between the total signal and the DNA fragments' total mass [96]. Although the normalization corrects for large amounts of signal drifts and scaling, we suspect that the rescaling is not perfect. Furthermore, a fraction of cells might not have been released from the synchronization or might have died in the process. To compensate for the reduced range of replication, we introduced a global background and a constant scaling factor for each time point as (genome-wide) parameters. Since these parameters are genome wide, they affect all origins simultaneously, and thus, the relationship between the local SM parameters $t_{1/2}$ and t_w is not significantly affected.

Data statistics. In performing the least-squares fit, we assume that the data points are normally distributed. This would be the case if we had many datasets to average over. However, only two independent and nominally equivalent time-course experiments are available. Subtracting one dataset from the other shows that the data points are actually not normally distributed (Appendix 3.A.1). Rather, the distribution has an exponential-like tail to the right. Instead of forming and using a more appropriate likelihood function, we used an exponential likelihood function to refit the data to test whether this likelihood function would change the parameters significantly. Our rationale is that since a normal distribution is too narrow and an exponential distribution too broad, the two possibilities bracket the true likelihood (see Fig. 3.12D in Appendix 3.A.1). We found that the robust fit leads to shifts in v by ≈ -0.2 kb/min, origin positions by $\approx \pm 1$ kb, $t_{1/2}$ by ≈ -3 min, and t_w by ≈ -2 min (Fig. 3.13). These small uncertainties do not affect our overall conclusions.

S-phase duration. In the microarray experiment, the progress of replication is monitored by flow cytometry [96]. The flow-cytometry data show that DNA content stops increasing after 60±10 minutes into S phase (Fig. 3.10A). We therefore estimate S phase to be 60 min. With our definition of potential efficiency in terms of $\Phi(t_{\text{end}})$, a change in t_{end} changes the potential efficiency of every origin. (Potential efficiencies as a function of t_{end} can be estimated from Fig. 3.3B.) Still, the trend that later-firing origins have lower potential efficiency remains valid, as is the trend between utilized and potential efficiency shown in Fig. 3.6C.

3.4.3 Summary of limitations

We summarize the effects of the above mentioned limitations:

1. There is starting-time asynchrony in the cell population probed. We extended the formulation to account for such asynchrony and found it to be consistent with a normal distribution with standard deviation $= 2$ min. Refitting the SM to part of the microarray data using the estimated asynchrony, we found that v shifted by $\approx$ -0.3 kb/min, $t_{1/2}$ by $\approx$ -0.5 min, and t_w by $\approx$ -1 min (Fig. 3.9).

2. The microarray dataset analyzed has a spatial resolution of $\approx$ 2 kb. This limits us to resolve origins that are closer than 5 kb apart.

3. The data do not cover the entire range of replication fraction (0 to 100%), perhaps because of contamination and imperfect signal normalization. We deal with these

artifacts by introducing a scaling factor for each time point and a background as fitting parameters (see Sec. 3.2).

4. All the fit parameters have a small systematic uncertainty that originates from an incomplete knowledge of the likelihood function (Figs. 3.11 and 3.12). We found that using an alternative form of the likelihood function shifts v by ≈ -0.2 kb/min*, origin positions by $\approx \pm 1$ kb, $t_{1/2}$ by ≈ -3 min, and t_w by ≈ -2 min.

5. From the flow-cytometry data, we estimated S phase to be 60±10 min. Uncertainty in the length of S phase affects the values of the efficiencies extracted but not the relationships shown in Fig. 3.6C and D.

In summary, artifacts in the microarray data result in small uncertainties in the absolute values of the extracted parameters but do not significantly alter our findings. In particular, the relationship that $t_{1/2} \approx t_w$ remains valid. The $t_{1/2}$-t_w trend clearly reveals that initiation times and the precision of timing are correlated. This relationship has also been observed by recent analyses [42, 23] and contrasts with the view that origin initiation is timed in a nearly deterministic fashion. We continue the discussion of this relationship in Chapter 4.

3.A Appendix

3.A.1 Statistical details of the fits

Here, we discuss in more detail the various fits described in the main text. We start by recalling the definition of the χ^2 statistic:

$$\chi^2 = \sum_{i=1}^{N_d} \frac{(f_i - d_i)^2}{\sigma_i^2} \,, \tag{3.5}$$

where N_d is the number of data points, f_i is the model value, d_i is the data (measurement) value, and σ_i is the standard deviation of the measurement d_i. The use of χ^2 statistic (least-squares fitting) asserts that statistical fluctuations affect each data point d_i independently

*Correcting for starting-time asynchrony and likelihood function makes the fork velocity closer to the one reported in [48].

t (min)	10	15	20	25	30	35	40	45
σ_t	1.16	1.43	1.96	2.46	2.96	3.37	3.05	3.00

Table 3.1: Standard deviation of single-measurement noise for time-course microarray. Top row: t is the time at which the measurement is made. Bottom row: σ_t is the standard deviation of the measurement noise for time point t.

and that the fluctuations are normally distributed, with mean 0 and standard deviation σ_i, which we denote N(0,σ_i). As we shall see, a detailed examination of the fluctuations shows that the assumptions for least-squares fits are not strictly met.

Ideally, the noise distribution for each data point would be estimated by repeating the experiment a large number of times. McCune *et al.* repeated their experiment once, meaning that there are just two measurements of each data point [57]. To examine the distribution of fluctuations, we consider the distribution of the differences between the timing curves from both experiments, calculated data point by data point (Fig. 3.11A). A cursory examination shows that the fluctuations vary notably with time: earlier time points show smaller fluctuations than later ones. We thus grouped the fluctuations by time points. Within each time point, fluctuations are homogeneous, except for an obvious upward bias corresponding to the data points representing Chromosome I (Fig. 3.11B). We observed a similar bias from Chromosome I in all 8 time points and thus excluded the data from the set of residuals used to estimate the distribution of fluctuations. (However, we did not exclude Chromosome I from our model fits.)

Excluding the differences from Chromosome I, we compiled histograms for the 8 time points (Fig. 3.12C). These histograms estimate probability distribution functions for the differences between two noisy measurements. For curve fitting, we need to estimate the distribution of a *single* noisy measurement. Elementary properties of the variance imply that, for two independent random variables X and Y, Var[$X - Y$] = Var[X] + Var[Y]. For two independently and identically distributed random variables, the standard deviations of the differences are then $\sqrt{2}$ times larger than the standard deviation for single-measurement noise. Correcting for this factor, we found the standard deviations recorded in Tab. 3.1. (The caption to Fig. 3.12C gives the uncorrected values.)

To examine the fluctuation distributions further, we rescaled the fluctuations for each data point by dividing by the standard deviation for that time point. After rescaling, all

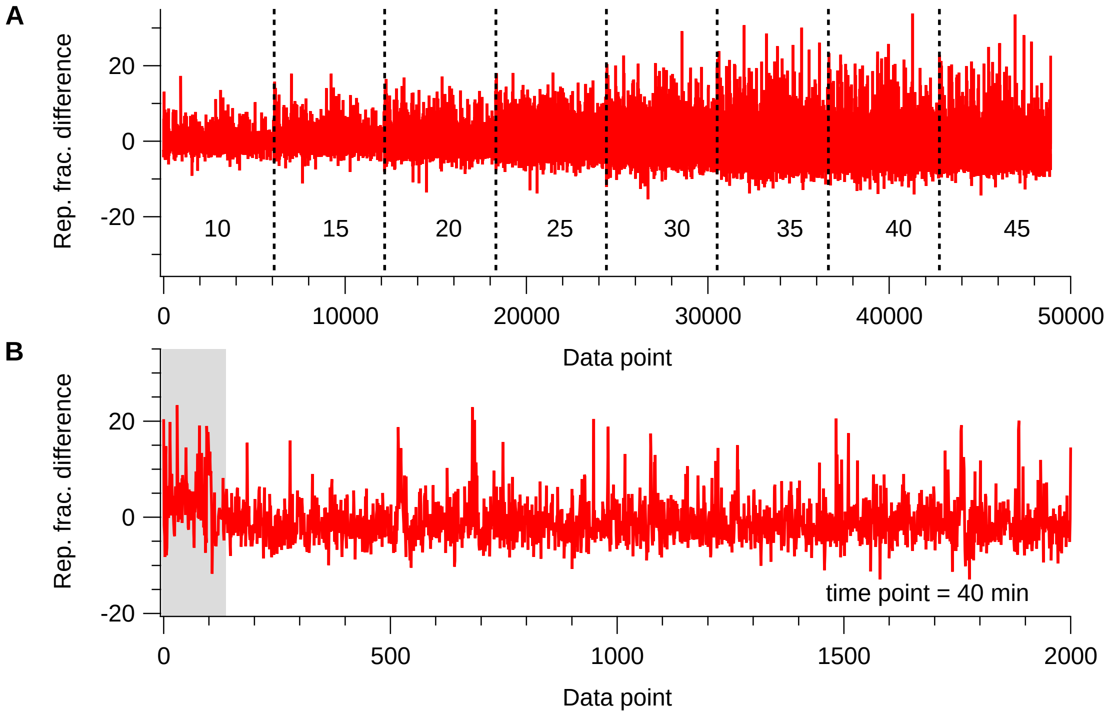


Figure 3.11: Statistics of the difference between two equivalent experiments, part 1. **A**. Difference between two equivalent experiments from [57]. The differences between the replication fraction of two nominally equivalent experiment are shown serially in time. The fluctuation of the differences varies across different time points. **B**. Differences for the first 2000 data points of time-point 40 are shown. The upward bias in the shaded region corresponds to Chromosome I. All time points have this bias.

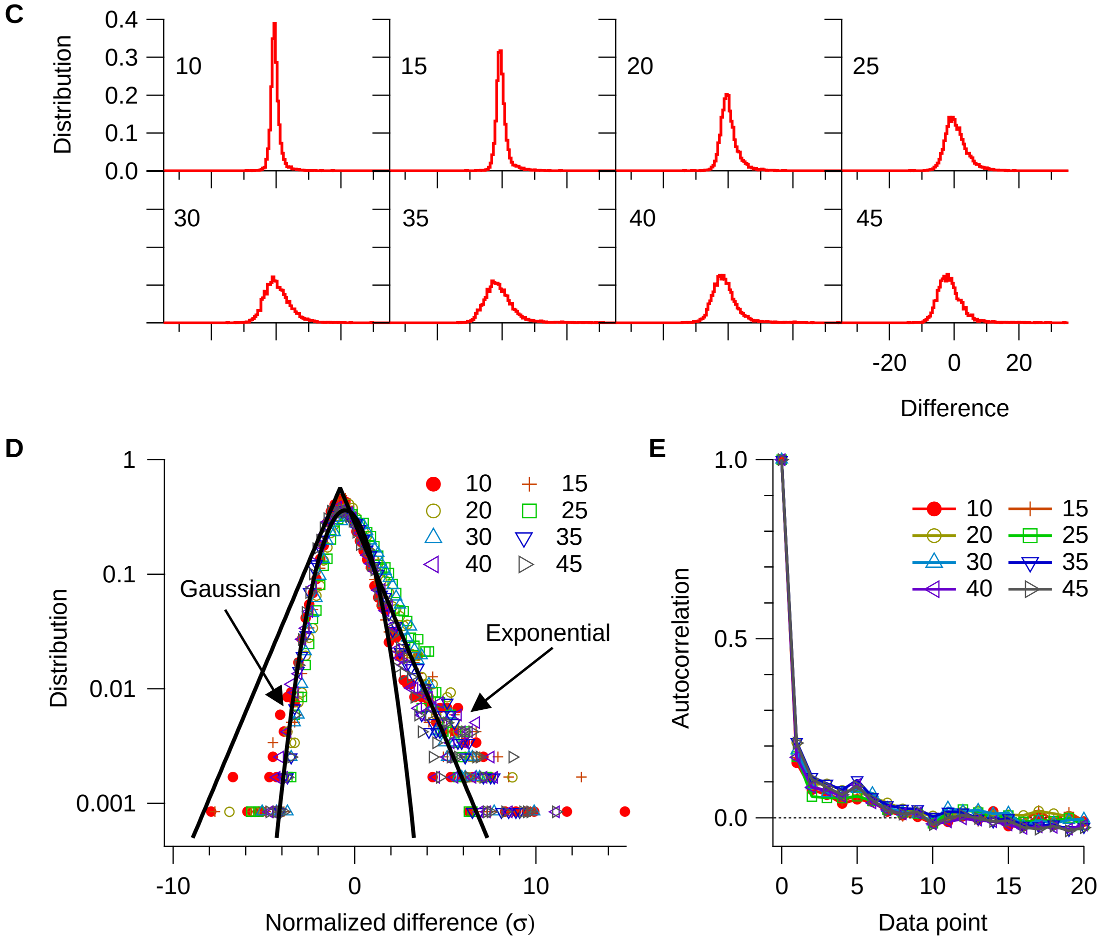


Figure 3.12: Statistics of the difference between two equivalent experiments, part 2. **C**. Histograms of the differences for different time points. In making the the histograms (bin width = 0.5), we excluded the first 200 data points of each time point because of the apparent upward bias. These data points correspond almost exactly to Chromosome I. The standard deviation of the differences (in sequence of increasing time points) are 1.64, 2.03, 2.77, 3.48, 4.18. 4.76. 4.31, and 4.25. The σ_t values used in the fits equal these standard deviations divided by $\sqrt{2}$. **D**. The histograms (bin width = 0.2) in C collapse onto the same distribution after scaling the differences for each time point with its corresponding σ_t. Small deviations are Gaussian like, while large positive deviations are exponential. **E**. Autocorrelation of the differences. The autocorrelation shown excludes the first 200 data points, as well.

eight histograms collapse to a single distribution (Fig. 3.12D). This confirms that the noise fluctuations depend only on a reduced variable $(f_i - d_i)/\sigma_i$, as assumed when writing down Eq. 3.5. Unfortunately, two problems need to be addressed in order to perform a rigorous fit. First, the distributions are not normally distributed (Fig. 3.12C). In particular, the positive-valued tails are approximately exponential, implying that large fluctuations are much more likely than those suggested by a noise model (likelihood function) based on Gaussian statistics. Second, the distribution is clearly skewed (asymmetric about 0). This means that the noise distributions of the two experiments are not identical and that the σ values obtained from the $\sqrt{2}$ scaling might not approximate the deviation of the single-measurement noise well. (It is easy to prove that the difference between two independently and identically distributed random variables must be distributed symmetrically about zero.) Without further measurements, it is difficult to infer the actual form of the noise distribution. One further test examines the independence of fluctuations in one data point compared to another. We checked this by computing the autocorrelation function of the (scaled) residuals. The autocorrelation curves collapse, and there is only a weak correlation in the first few data points (Fig. 3.12E). Thus, the assumption of independence is reasonably well satisfied.

At this point, we have established that it is reasonable to treat the fluctuations in each data point separately and that the fluctuations are a function only of the reduced variable $(f_i - d_i)/\sigma_i$. Although we do not know the exact form of the likelihood function, we can examine how sensitive our model fits are to its precise form. Thus, we compared least-squares fits (assume Gaussian likelihood function) with robust fits (assume Exponential likelihood and use $\chi^2 = \sum_{i=1}^{N_d} |f_i - d_i|/\sigma_i$). To compare the fits, we fit to Chromosome XI, and the results are shown in Fig. 3.13. In general, we found little to distinguish between the results of the two fits. The main difference is that there are systematic shifts between corresponding parameters. The robust fit shifts the global v by ≈ -0.2 kb/min, origin positions by $\approx \pm 1$ kb, $t_{1/2}$ by ≈ -3 min, and t_w by ≈ -2 min. We speculate that using the actual noise distribution to fit would give parameters whose values are in between those obtained from a least-squares fit and those obtained from a robust fit.

Since least-squares and robust fits give similar parameter values, we decided to adopt the more standard least-squares χ^2 statistic. We keep in mind that any resulting P values will be severely underestimated, as they fail to account for the exponential tail of the distribution. For a similar reason, the statistical errors for the parameters estimated by the

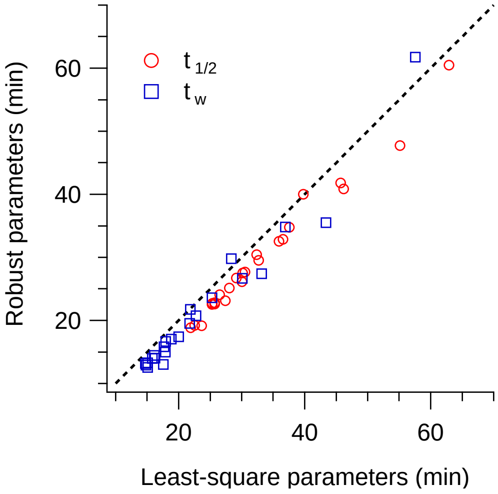


Figure 3.13: Comparison between least-squares and robust fit parameters for Chromosome XI. The x-axis corresponds to the least-squares fit, and the y-axis to the robust fit. Dotted line shows $y = x$. The least-squares $t_{1/2}$ (t_w) values are on average 3.24 (0.73) min larger than the robust $t_{1/2}$ (t_w) values.

fit will be underestimated. In reporting our fits, instead of using χ^2, we follow common practice and record the "reduced chi square" $\chi^2_\nu \equiv \chi^2/\nu$, where ν is the number of degrees of freedom ($\nu = N_d - N_p$, with N_d the number of data points and N_p the number of free parameters in the fit). For $\nu \gg 1$, which is always true in our analysis, the χ^2_ν statistic is expected to be distributed as N$(1, \sqrt{2/\nu})$. Again, we note that the exponential tail of the noise fluctuations will increase the expected standard deviation of the χ^2_ν statistic significantly.

Before proceeding to whole-genome fits, we first made a detailed comparison of the variable-fork-velocity model (VVSM), SM, and MIM on Chromosome XI*, which has $N_d = 2678$ and $N_p = 99$, 76, and 54 for the VVSM, SM, and MIM, respectively. The χ^2_ν values for the three models are 2.29, 2.48, and 2.76. These values exceed the expected χ^2_ν value of 1 by 42, 53, and 63 standard deviations. Given the uncertainty in the distribution of χ^2_ν, we did not reject the fits but attempted a more qualitative description of the fit quality (Fig. 3.14). The fit residuals and their distributions are all quite similar (Fig. 3.14A and B). The autocorrelation functions decrease to zero within a few data points (Fig. 3.14C), suggesting that the fits do capture most of the details of the data. The similarity in the results of

*The multiple-initiator model (MIM) is a model introduced in Chapter 4.

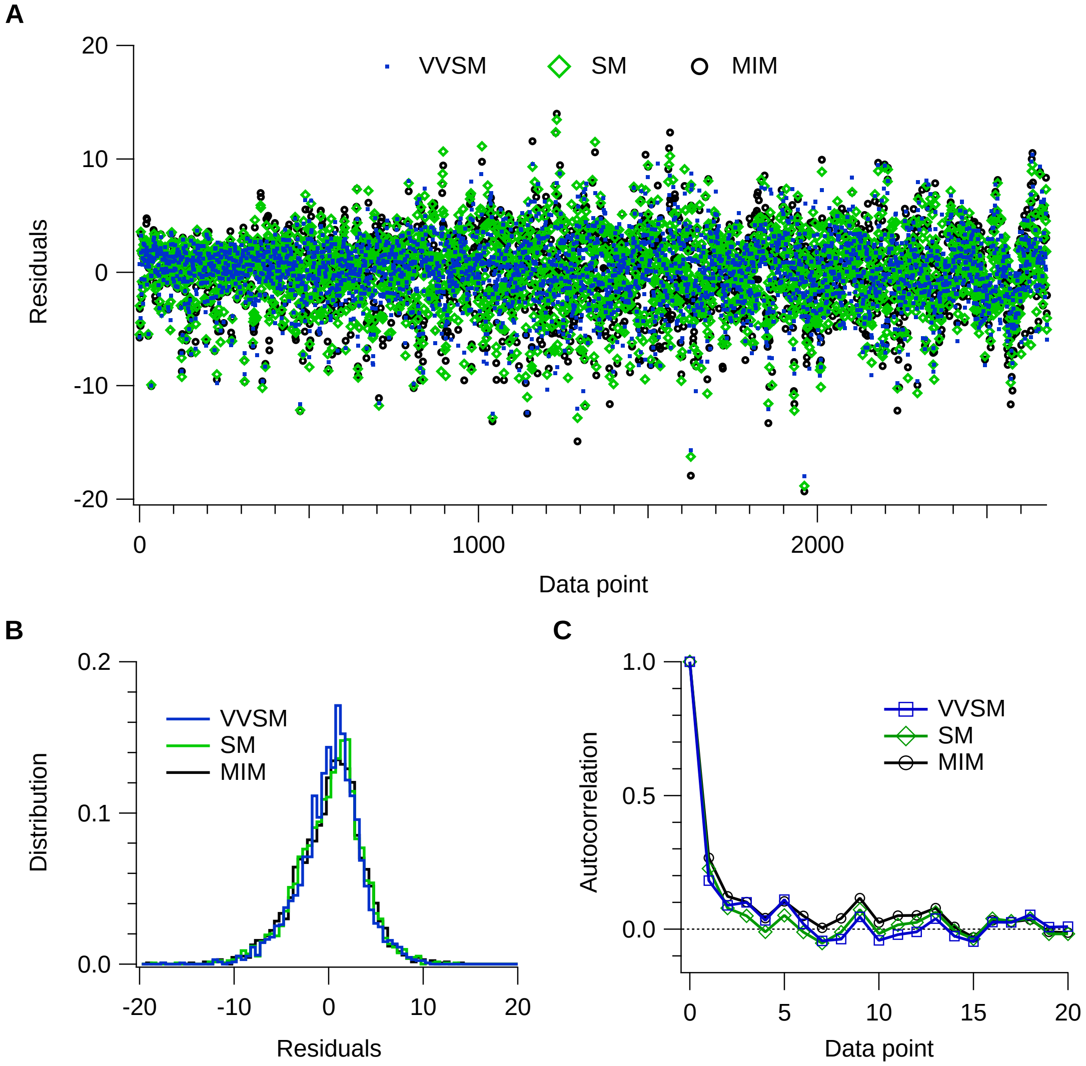


Figure 3.14: Residuals of different model fits. **A**. Residuals of the model fits to Chromosome XI. Markers correspond to the residuals of the three different model fits, VVSM, SM, and MIM, discussed in the text. The residuals are plotted serially in time points. **B**. Histogram of the residuals with bin width = 0.5. The standard deviations of the VVSM, SM, and MIM residuals are 3.21, 3.42, and 3.52, respectively **C**. Autocorrelation of the residuals of the VVSM, SM, and MIM fits.

the three models justifies favoring the model with fewest parameters (MIM model). Repeating the comparison for whole-genome fits, we found χ^2_ν for the SM and MIM genome-wide fits: 4.91 and 5.83 (ν = 48129 and 48481).

3.A.2 Variable-velocity sigmoid model

The formalism introduced in Sec. 2.1 can be extended to incorporate a space-time-dependent fork velocity $v(x, t)$. We generated a spatially varying velocity $v(x)$ as follows: The summand in Eq. 2.5 in the main text is non-zero only when Δx_p contains an origin at x_i, implying that the sum is really only over $p = i$. By replacing the global v by a local v_i, we associated a different fork velocity with each origin. In this way, we obtained spatially varying fork velocities. Since the origins are well localized, this scenario roughly corresponds to velocity variation in zones that are $\approx$ 50 kb in size. Generalizing further, with a variable fork velocity $v_i(t)$, the edges of the triangle in Fig. 2.1 would be curved. The goal is then to find the time along the curved edge by solving

$$\int_{t_e}^{t} v_i(t)dt = |x - x_p| \tag{3.6}$$

for t_e. Here, t_e is a function of t, $|x - x_p|$, and the parameters that form $v(t)$. This generalizes the constant-velocity case, where $t_e = t - |x - x_p|/v$. Replacing the argument $t - |x - x_p|/v$ used previously with $t_e(t, |x - x_p|, v_{i,\cdots})$ [with $v_{i,\cdots}$ representing the parameters that describe $v_i(t)$], one obtains a formalism that allows for a time-dependent fork velocity. In the fits, we kept the velocity constant in time. This assumption is consistent with independent evidence that the velocity is constant throughout S phase [106].

We used this "variable-velocity-sigmoid model" (VVSM), the SM, and the MIM to fit Chromosome XI (Fig. 3.15). Each of the three models captures most of the variations in the data, explaining 98.87% (VVSM), 98.77% (SM), and 98.62% (MIM) of the variance of the raw data. We also showed that the distributions of the residuals of the three fits are very similar (Fig. 3.14B), indicating that the quality of the three fits is similar. Thus, we conclude that constant-velocity models describe the replication kinetics as well as variable-velocity models.

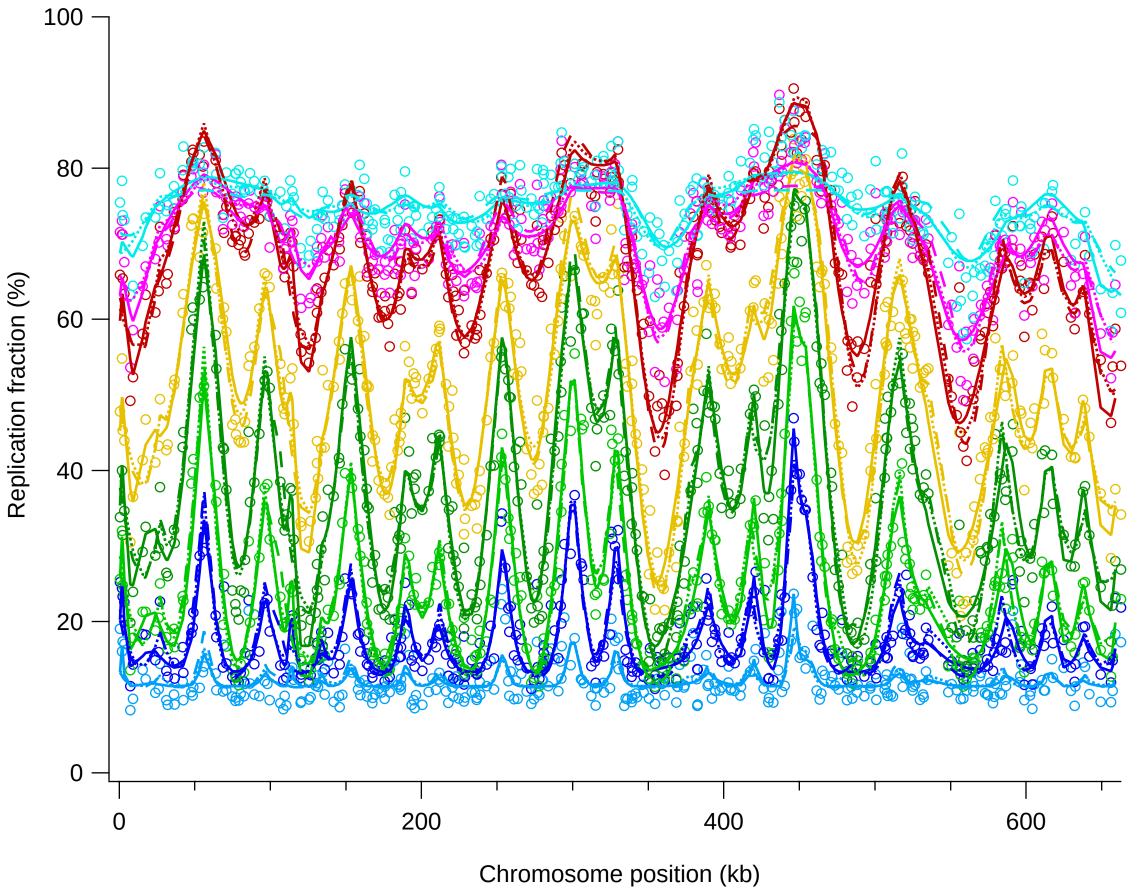


Figure 3.15: Fits to Chromosome XI. Markers are data; solid lines are fits from VVSM; dotted lines are fits from SM; and dashed lines are fits from MIM. The eight curves from bottom to top correspond to the replication fraction $f(x)$ at 10, 15, 20, 25, 30, 35, 40 and 45 min after release from the restriction temperature of *cdc*7-1. The dataset covers the genome at $\approx$ 2-kb resolution.

3.A.3 Fits to raw and smoothed data

It is common practice to analyze a smoothed version of microarray data so that peaks can be more easily identified. It is thus tempting to use smoothed data for curve fitting, as well. However, there are reasons to prefer fits to the raw, unsmoothed data. First, as a matter of principle, smoothing can only reduce the information available in a dataset and can never add to it. Second, the smoothing procedure correlates the statistical fluctuations among nearby data points, requiring a modification of standard least-squares fitting algorithms.

To test whether there are significant differences between the results of fitting to raw and to smoothed datasets, we repeated the SM fit of Chromosome XI using the smoothed data of [57]. The residuals (Fig. 3.16A) and their autocorrelation function (Fig. 3.16B) show a correlation among neighbouring points that results from the smoothing operation. The χ^2 statistic of standard least-squares routines then needs to be modified to explicitly account for the correlations [107]. In this case, the use of the standard statistic (Eq. 3.5) can bias the resulting parameter values. With this particular dataset, we found little practical difference between fitting to the raw and smoothed data. Both fits produced parameter values that mostly agreed to within 10%; only a few parameters, which correspond to less-apparent peaks in the microarray data, do not match well (Fig. 3.16C). Thus, it is unlikely that any substantive conclusions reached about this particular dataset would have been affected had we fit to the smoothed data using the standard χ^2 statistic; however, since it is just as easy to fit to raw data, we recommend doing this and encourage experimental groups to publish and make available the raw datasets.

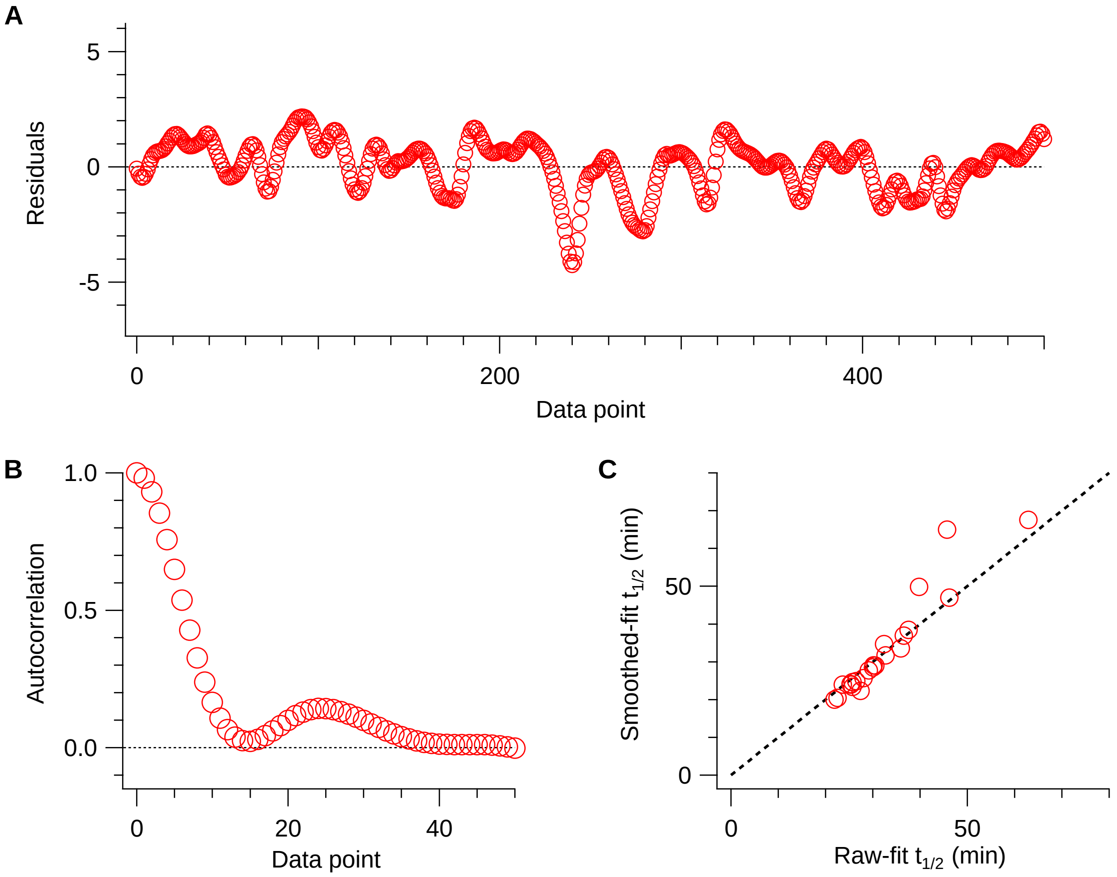


Figure 3.16: Fitting smoothed data: residual and parameter. **A**. Residuals of SM fit to the smoothed data of Chromosome XI. the first 500 of the 5136 data points of residuals are shown for clarity. The number of data points here is larger than that of the raw data (2678) because the smoothed data were also interpolated [57, 20]. **B**. Autocorrelation of residuals, showing the correlation in noise produced by the smoothing algorithm. **C**. Comparison of $t_{1/2}$ fit parameters. The x-axis corresponds to the raw-data fit, the y-axis to the smoothed-data fit. Dotted line shows $y = x$.

Chapter 4

Control of Origin Timing in Budding Yeast

In Chapter 3, we showed that the timing ($t_{1/2}$) and precision (t_w) of origin initiation are correlated in budding yeast. This correlation implies that earlier-firing origins initiate with less temporal variation and are potentially more efficient. Many of these highly efficient origins are not normally used because of passive replication but can help safeguard against replication stress such as fork stalls. What kind of mechanism can lead to this correlation? In this chapter, we propose a model, the multiple-initiator model (MIM), where earlier-firing origins have more initiator complexes loaded and are located at regions that have more-accessible chromatin structure. We refit the time-course microarray data with the MIM and show that the MIM captures the $t_{1/2}$-t_w trend qualitatively.

The model demonstrates how initiation can be stochastic and yet occur on average at defined times during S phase without an explicit time-measuring mechanism. Furthermore, we hypothesize that the initiators in this model correspond biologically to loaded pairs of minichromosome maintenance (MCM) complexes. We compare the initiator number extracted from the fit with independent measurement of MCM occupancy and show that the two correlate. We also investigate the correlation between the initiator number and two quantities that relate to chromatin accessibility, namely histone acetylation and deacetylation. This model is the first to suggest a detailed, testable, biochemically plausible mechanism for the regulation of replication timing in budding yeast. Because the elements of the model are found generally in eukaryotic organisms, one expects that it may apply to a broad range of eukaryotes. In terms of the three themes mentioned in Sec. 1.4, this chapter explains the biological significance of the model-extracted quantities and thus advances our understanding about the temporal aspect of replication in eukaryotes.

4.1 Introduction

The conventional understanding of origin timing control has two steps: first, early-or-late "time stamps" are formed on the origin loci; second, activators recognize specific time stamps at preprogrammed times to execute the timing program [94]. There are many candidates for the first step. Proposed time stamps include the level of histone acetylation [94], the openness of chromatin structure [95], and the localization of the origin loci within the nucleus [94, 71]. The idea that activators function at preprogrammed times for the second step is plausible, as there are factors whose activities are known to be modulated in different phases of the cell cycle. For instance, the CDK activity mentioned in Sec. 1.1 is low in G1 phase to allow licensing of potential origins but increases in S phase to prevent relicensing and re-replication [4]. On the other hand, the mechanisms by which the early/late activators recognize only early/late time stamps are elusive.

The results from our Sigmoid Model (SM) fit to time-course microarray data are incompatible with the above picture in two aspects. First, since the SM allows for any combination of median firing time ($t_{1/2}$) and time width (t_w), the model is capable of capturing a broad range of dynamics. In particular, the SM would result in distinct classes of early and late origins if the data were to strongly suggest that scenario. However, the fit reveals that the initiation timing is a continuum (Fig. 3.8) and supports the view that stochastic effects overwhelm the early-late distinction at the single-cell level. Second, the fit shows a tight and smooth correlation between $t_{1/2}$ and t_w (Fig. 3.4). If time-stamp-specific activators were functional at preprogrammed times, one would expect no correlation between $t_{1/2}$ and t_w and would expect t_w to be roughly the same for both early and late origins. Overall, our results support a scenario where origins do not form distinct classes and activators are not time-stamp specific. Our proposal is that initiation timing is controlled with the multiplicity of stochastic initiators and non-specific, random activators. In the presentation below, we first describe the technical details of the model in Sec. 4.2. Then, we investigate its biological significance in Sec. 4.3.

4.2 The multiple-initiator model (MIM)

Replication initiates at origins because there are initiator proteins bound to them. At some point during S phase, trans-acting activator proteins activate the initiators to start the unwinding and elongation. Suppose the dynamics between every activator and initiator were the same. Then, every initiator would fire according to a global firing-time distribution $\phi_o(t)$, and origins that have more initiators loaded would be more efficient. This situation arises because when multiple initiators bind near an origin site, it is always the earliest firing that counts. Other initiators cannot fire to re-replicate the same site [5]. One can assign an effective firing-timing distribution $\phi_{\text{eff}}(t, n)$ to the initiator cluster at an origin, with n being the number of initiators in the cluster. This distribution of the earliest firing times shows the same trend as the curves extracted from the microarray data: earlier-firing origins have narrower distributions (Compare Figs. 4.1B with 3.3A). We call this the "multiple-initiator model" (MIM), since the number of initiators determines the shape of the firing-time distribution. Note that because the MIM captures the $t_{1/2}$-t_w trend, all the properties of SM origins are captured (see Sec. 3.3.8).

For moderately large n ($n \gtrsim 10$), the selection of the first initiation among many causes $\phi_{\text{eff}}(t, n)$ to tend to a universal distribution, the Weibull distribution, regardless of the "details" of the $\phi_o(t)$ used [82]. As an example, the shape of $\phi_{\text{eff}}(t, n)$ would differ between using an increasing and a decreasing $\phi_o(t)$ but would not alter significantly between a linearly and a quadratically increasing $\phi_o(t)$. This robustness is an advantage of the model because it obviates the need for an accurate form of $\phi_o(t)$. The cumulative Weibull distribution has the form

$$\Phi_{\text{W}}(t) = 1 - e^{-\left(\frac{t}{\lambda}\right)^k}, \tag{4.1}$$

where λ is the scale factor and k the shape factor. If $\Phi_{\text{eff}}(t, n)$, the effective cumulative firing-time distribution, tends to $\Phi_{\text{W}}(t)$, Eq. 4.1 suggests that the plot of $-\ln[1 - \Phi_{\text{eff}}(t, n)]$ vs. t on a log-log scale for different n should be straight lines. The inset of Fig. 4.1B shows that this is indeed the case. We will explain why the lines have the same slope in the next section.

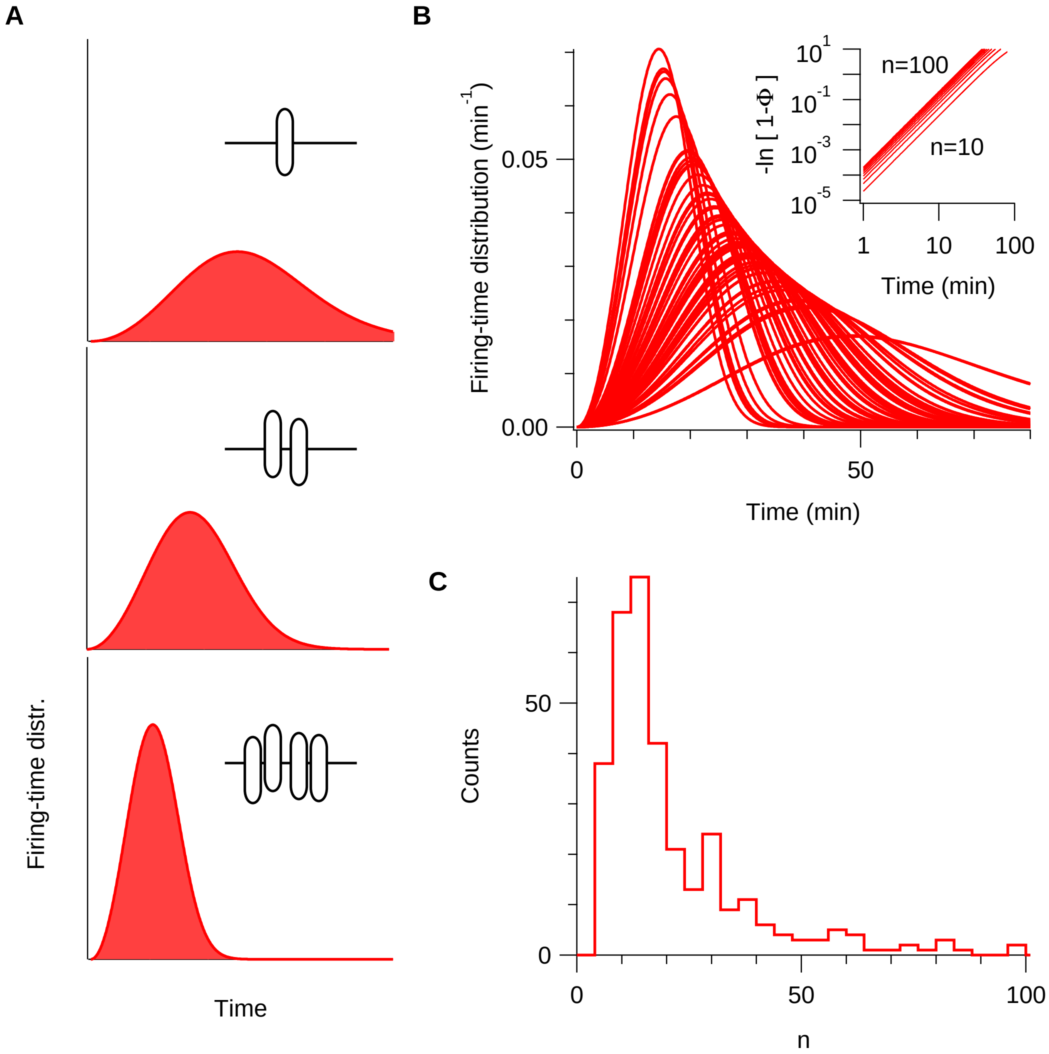


Figure 4.1: **A**. Illustration of the multiple-initiator model. Initiators are loaded onto an origin site. The effective firing-time distribution of the origin narrows and shifts to earlier times with increasing number of initiators. **B**. Firing-time distributions in the multiple-initiator model. The parameter n ranges from 2.2 to 96.4. These firing-time distributions resemble those shown in Fig. 3.3A. Inset shows $-\ln[1-\Phi_{\text{eff}}(t,n)]$ vs. t on log-log scale. The n in the subgraph goes from 10 to 100 in increments of 10. **C**. Histogram of n with bin width = 4. Values of n range from 2.2 to 96.4. Median value = 13.3; standard deviation = 16.7.

4.2.1 Fit parameters for the MIM

The MIM, like the SM, is parameterized with a constant fork velocity (global in space and time), a constant background (global in space and time), eight constant normalization factors (global in space and one for each time point), and 732 starting origin positions (see Sec. 3.2). The two firing-time-distribution parameters $t_{1/2}$ and t_w for each origin in the SM are replaced by a single parameter n in the MIM. As we will discuss below, n depends on the number of initiator molecules but is modified by effects such as chromatin accessibility. There are two additional parameters, $t^*_{1/2}$ and r^*, that describe the global firing-time distribution of all initiators (Eq. 4.3). These are absent from the SM, as the SM does not assume that the firing-time distributions are related. All local parameters of the MIM are tabulated in Supplementary Table II and all global parameters in Supplementary Table III (Appendix 4.A.2).

We again use the Hill equation to describe the global cumulative distribution $\Phi_o(t)$. Using Eq. 2.14, the $\Phi_{\text{eff}}(t, n)$ for a cluster of n initiators is

$$\Phi_{\text{eff}}(t, n) = 1 - [1 - \Phi_o(t)]^n \,, \tag{4.2}$$

where

$$\Phi_o(t) = \frac{1}{1 + \left(\frac{t^*_{1/2}}{t}\right)^{r^*}} . \tag{4.3}$$

The quantity $t^*_{1/2}$ is the median time of the firing-time distribution for a single initiator, and r^* is the distribution's rate of increase (see Appendix 4.A.1 for a discussion on the use of Hill equation). Mathematically, the MIM reveals an interesting feature when expressed in terms of initiation rates. Equation 2.10 suggests that the global cumulative distribution can be rewritten as $\Phi_o(t) = 1 - \exp[-\int_0^t I_o(t')dt']$, where $I_o(t)$ is the corresponding global initiation rate that every initiator follows. Substituting this into Eq. 4.2, one obtains

$$\Phi_{\text{eff}}(t, n) = 1 - \exp\left[-n \int_0^t I_o(t')dt'\right] . \tag{4.4}$$

Eq. 4.4 shows that the MIM includes a class of models where the initiation rate of each origin is described by a scaling constant n times the global initiation rate $I_o(t)$. This convenient feature allows one to easily identify whether a model falls into the MIM class. This feature also explains why the lines have the same slope in the inset of Fig. 4.1B. Comparing Eqs. 4.4 with 4.1, we see that the slope is determined by $I_o(t)$ and thus universal, while the intercept is determined by n and shifts upward with increasing n.

Due to the different form of firing-time distribution used in the MIM, the criteria for origin elimination are slightly modified from the SM ones (see Sec. 3.2). The MIM predicts strictly that earlier-firing origins are more efficient; thus, no contamination effects are modelled by the MIM (see discussion in **Origin position** under Sec. 3.2). We then eliminated origins via a single criterion of $\Phi_i(t = 60 \text{ min}) < 0.4$. The change is consistent with the observation that the firing-time distribution in Eq. 4.2 tends to result in more efficient origins than the Hill function does. Similar to the SM fits, individual-chromosome fits were done first before performing the genome-wide fit.

After eliminating origins with the above criterion, the MIM gave 337 origins. Of the 337, 234 colocalize with the 275 origins identified in [96]. Of the remaining 103 origins, 70% colocalize with known origins from the OriDB database [97] within 5 kb. Together, the SM and MIM gave 357 distinct origins: 322 are in both SM and MIM, 20 in only SM, and 15 in only MIM. Among these, 116 were not identified in [96], and 71% of them colocalize with known origins from the OriDB database within 5 kb. As is for the SM, the parameter values of the MIM origins are not significantly affected by the experimental limitations (Sec. 3.4). The average error on the origin position is 1.1 kb for the MIM. Inefficient origins have large errors ($n\pm$ 30%) but do not alter the replication kinetics. Uncertainty in the fit likelihood roughly translates to a ± 1 uncertainty in n. Overall, the imperfections of the data have minor effects on the MIM.

Results of the MIM fits are shown in Fig. 4.8. In the SM, each origin has three parameters: x, $t_{1/2}$, and t_w. In the MIM, each origin has only two: x and n. Although the number of parameters used decreased by nearly 1/3, the MIM fits are similar to the SM fits. This suggests that the MIM is likely the more appropriate model for replication in budding yeast (Appendices 3.A.1 and 3.A.2).

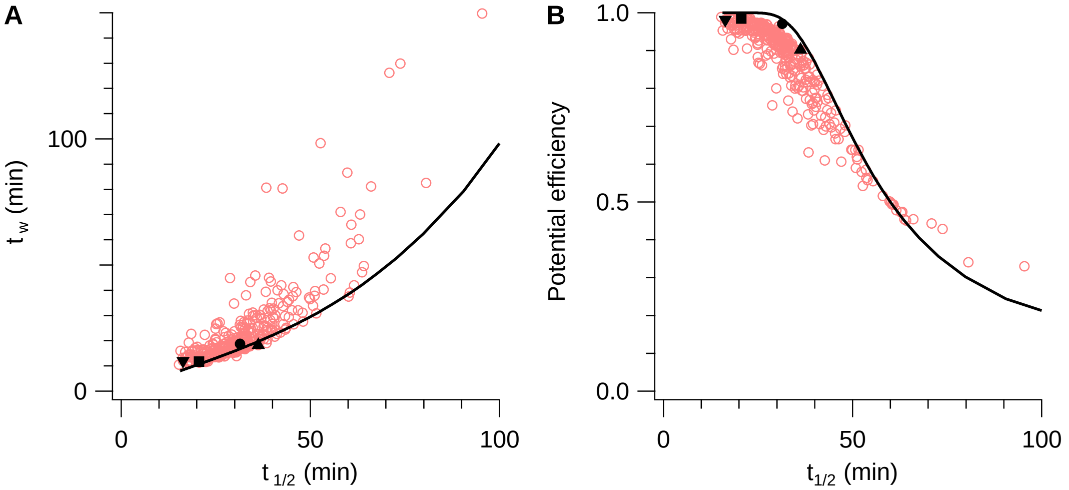


Figure 4.2: MIM matches the $t_{1/2}$-vs.-t_w and the $t_{1/2}$-vs.-efficiency trend. **A**. Markers are results from the SM, a reproduction of Fig. 3.4. The curve is the result from the multiple-initiator model (MIM) using $\Phi_o(t) = 1/[1 + (76/t)^3]$. The parameter values 76 min and 3 result from the the genome-wide MIM fit. Solid (darker) square, triangle, inverted triangle, and circle correspond to ARS 413, 501, 606, and 1114.5, respectively. **B**. Potential efficiency vs. $t_{1/2}$. Markers are results from SM, taken from Fig. 3.6D. The solid curve is calculated using the parameters obtained from the genome-wide MIM fit. The solid, darker markers are as in **A**.

4.3 Significance of the MIM

4.3.1 Link between $t_{1/2}$ and t_w

The SM describes the replication program with a sigmoid at each origin, in effect giving each origin an independent firing program. The correlation between the SM parameters $t_{1/2}$ and t_w, however, suggests that the firing programs of all origins are linked. The MIM provides the link: while each origin consists of a particular number of initiators, *all initiators are the same* (meaning all are described by the same $\phi_o(t)$). The similarity between Figs. 4.1B and 3.3A is evidence that the MIM scenario is promising. The MIM also captures qualitatively the $t_{1/2}$-vs.-t_w trend and the relationship between $t_{1/2}$ and potential efficiency (Fig. 4.2). We emphasize that these similarities are biologically significant because the MIM eliminates the need for time-measuring activators and shows that a robust timing order can be built from indistinguishable, stochastic initiators and activators.

One observation that has been used to support the view of time-specific activators in *S. cerevisiae* is that the later-replicating regions of the genome suffer significant delay in a *clb5*Δ strain compared to that in the wild type, while the earlier-replicating regions are largely unaffected [57]. The interpretation offered was that the Clb5 proteins specifically help activate origins that are marked late; thus, upon decreasing their abundance in the mutant, only the late origins are inhibited. The MIM offers another interpretation: The decrease in Clb5 proteins suppresses the number or function of non-specific activators and affects all origins. Scaling the activator number N_a by a factor $\alpha < 1$ in the presence of n initiators is equivalent to scaling the initiator number n by α in the presence of N_a activators. Under such a scenario, how would the firing time change as a function of n?

By using Eqs. 4.2 and 4.3 and setting $\Phi_{\text{eff}}(t_{1/2}, n) = 1/2$, we obtained $t_{1/2}$ as a function of n:

$$t_{1/2}(n) = t^*_{1/2} \left(\sqrt[n]{2} - 1 \right)^{\frac{1}{r^*}} , \tag{4.5}$$

where $t^*_{1/2}$ and r^* are the parameters that define $\Phi_o(t)$. For completeness, we also obtained $t_w = t_{3/4} - t_{1/4}$ as a function of n:

$$t_w(n) = t^*_{1/2} \left[\left(\sqrt[n]{4} - 1 \right)^{\frac{1}{r^*}} - \left(\sqrt[n]{4/3} - 1 \right)^{\frac{1}{r^*}} \right] . \tag{4.6}$$

The $t_{1/2}(n)$ for normal initiation in the wild type and delayed initiation in the *clb5*Δ mutant are plotted in Fig. 4.3A. In the *clb5*Δ mutant, the initiator number n is scaled down by α, and $t_{1/2}$ increases as a consequence. (We used $\alpha = 0.7$ as an illustration; any value less than 1 would give the same trend.) Figure 4.3B shows the difference in $t_{1/2}$ between the *clb5*Δ mutant and the wild type. The key observation is that origins with large n (earlier-firing origins) are affected much less than origins with small n (later-firing origins). This conclusion matches the observation reported in [57]—that later-replicating regions suffer significant delay, while earlier-replicating regions are largely unaffected—but does not need time-specific activators that target only late origins*. The result strengthens our proposal that initiation timing is controlled by the number of initiators and non-specific, random activators.

*In a future analysis, one can quantify the qualitative agreement by fitting the MIM to the $f(x,t)$ in *clb5*Δ. According to our reasoning, the wild-type n and *clb5*Δ n should be linearly proportional with a slope of α.

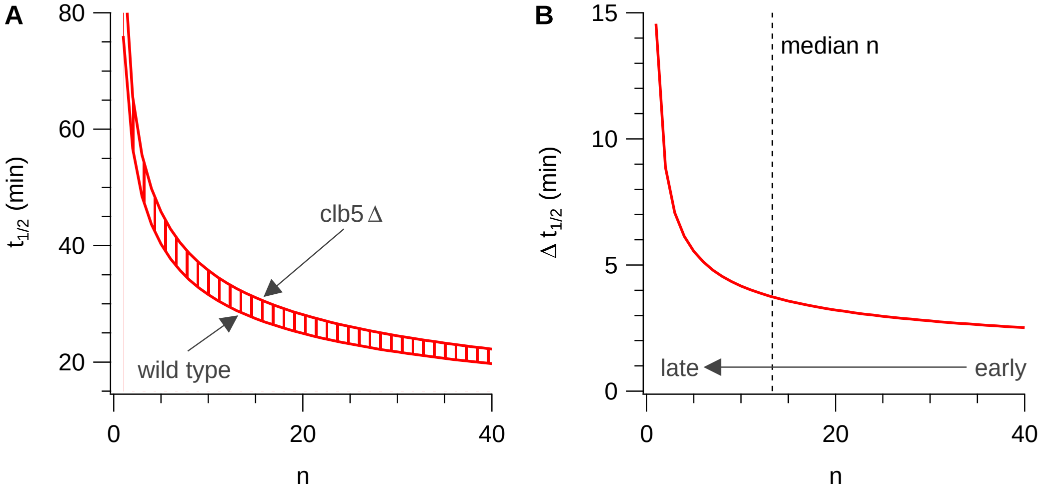


Figure 4.3: Relationship between $t_{1/2}$ and n in the MIM. **A**. The x-axis is the initiator number n. The y-axis is the $t_{1/2}$, calculated via Eq. 4.5 using $t^*_{1/2} = 76$ min and $r^* = 3$. The clb5Δ mutant is calculated by replacing n with αn. Here , α is chosen to be 0.7 for illustration. **B**. The y-axis is the difference between the two curves in A. The dashed line marks the median of the n distribution as deduced from Fig. 4.1C.

4.3.2 An initiator candidate: the MCM complex

We have shown that the MIM captures the origin initiation properties revealed by the SM. The MIM suggests that the origin's timing is controlled by stochastic initiators rather than deterministic time stamps. It also shows how controlled timing patterns form without time-measuring activators. In this section, we propose a biologically plausible candidate for the initiator: the minichromosome maintenance (MCM) complex [5]. MCM complexes are associated with the unwinding of DNA, one of the initial steps in origin activation, and are loaded in excess onto the DNA prior to S phase (Sec. 1.1) [99, 108, 109]. Below, we investigate the biological interpretations and implications of the MIM in light of MCM biology.

The parameter n for the number of initiators does not have to be an integer, as it represents the average number of initiators bound to an origin. Because the value of n is coupled to the $\phi_o(t)$ used, relative variations in n between origins are more significant than absolute values. From the data of McCune *et al.* in [57], we found a 43-fold range between the smallest and largest n (n =2.2 and 96.4, respectively), which is larger than one might expect for variation in MCM loading at a single origin. However, the majority of origins falls within a 20-fold range, between 2 and 40 (Fig. 4.1C), which is consistent with experimentally observed levels of MCM loading [108, 109].

In addition, other factors such as chromatin structure and origin-recognition-complex occupancy can influence the loading of initiators. Previous work has hypothesized that the time required for an activator to find an origin can limit the origin-initiation rate [37]. The search times in heterochromatin regions could be longer than in euchromatin regions because initiators in heterochromatin are less accessible to activators [110]. This in turn implies that initiation rates should be "scaled down" in heterochromatin regions (initiation rates are related to firing-time distributions via Eq. 2.12). For the same accessibility reasons, the loading of initiators may also be reduced in the heterochromatin regions. Put together, variations in chromatin structure and in the loading of initiators increase the range of inferred n values. Thus, the fit parameter n represents not simply the number of initiators but rather the combined effects of origin accessibility and initiator multiplicity. As an illustration, if the structure of chromatin identically affects initiation loading and origin activation, the number of initiators would be $\sqrt{n}$, and the range of initiator number will cover a 6.5-fold difference.

In the MIM, the parameter n has direct implications for the timing of the replication program. Figure 3.8A shows the histogram of all the $t_{1/2}$ values we extracted. The mode of the histogram suggests a typical firing time. The MIM implies that the typical firing time is related to the average number of initiators loaded onto the origin sites. The histogram also shows that no $t_{1/2}$ is earlier than 15 min. In the context of the MIM, this observation implies an upper limit to the number of origin-bound initiators, which corresponds to the largest n-value found (≈ 43 after normalizing the smallest n to 1). To judge whether such a value is reasonable, we make a crude estimate of the largest biologically plausible value of n, which can be associated with a close packing of the double MCM hexamers. Each pair of hexamers is roughly 30 nm long (0.1 kb) [111], and we imagine the pairs spreading out from an origin-recognition-complex (ORC) binding site while still closely packed around it. If the pairs were to spread more than ± 2.5 kb to either side of the ORC loading site, we would, in our analysis, assign more than one origin to that region. The largest value of n for a single origin is then $\approx$ 5 kb / 0.1 kb = 50, which is greater than the largest value found. In summary, the MIM is not only a mathematical model that captures the replication kinetics in budding yeast but also a plausible biochemical model that gives insight to the mechanism of initiation timing control.

4.3.3 Correlation with MCM complex

Having proposed MCM as the initiator, we now seek evidence for a correlation between the MCM number and the MIM n. Before discussing the experimental results, we briefly explain how three techniques, ChIP-PCR, ChIP-chip, and ChIP-seq, are used to probe MCM number. Common to all three is chromatin immunoprecipitation (ChIP), a technique used to isolate the DNA bound to specific proteins. The isolation procedures involve crosslinking DNA-bound proteins to DNA, fragmenting the DNA into sub-kb-sized pieces, precipitating out the desired protein by a specific antibody, and dissolving the cross links to isolate the DNA. In all three techniques, ChIP is used to isolate the MCM-bound DNA. The three techniques, however, differ in the way that the isolated DNA is detected:

> **ChIP-PCR**. Polymerase chain reaction (PCR) is a technique used to amplify the copy number of a particular sequence of DNA. From the pool of MCM-bound DNA fragments, one can choose particular sequences (e.g., those that correspond to early origins) to amplify using PCR. Since the number of DNA copies for a particular sequence reflects the number of MCM bound to that sequence, ChIP-PCR can be used to compare the relative MCM occupancy at a few different origin loci.
>
> **ChIP-chip**. Here, PCR is replaced by microarray chips that can measure the variations in DNA copy number across the genome. The result is a genome-wide profile of relative MCM occupancy. The ChIP-chip signal reported is usually relative to a control experiment, where the number of DNA copies measured is known to be constant. It has been observed that the finite dynamic range of microarray data, defined as the ratio between the signal and the control, can limit the signal; i.e., the actual ratio between the quantity and the control is smaller or larger than the microarray can reliably probe [112].
>
> **ChIP-seq**. Here, the isolated DNA is sequenced directly. ChIP-seq generates a genome-wide profile of relative MCM occupancy similar to ChIP-chip but provides larger dynamic range, higher resolution, and fewer normalization issues [64].

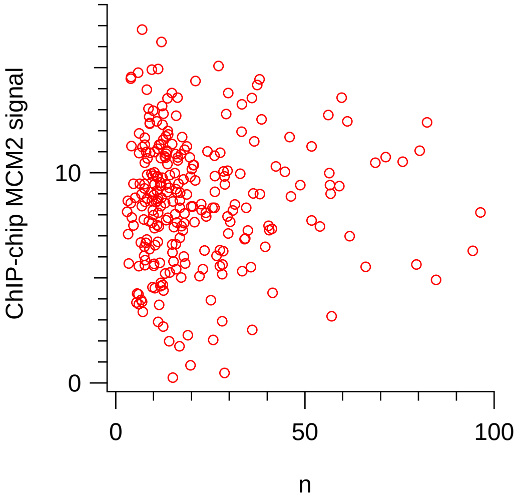


Figure 4.4: ChIP-chip signal of MCM occupancy vs. parameter n. The y-axis is the ChIP-chip signal for MCM2 (the first unit of the MCM2-7 hexamer ring) occupancy from [113]; the x-axis is the parameter n that we extracted from the MIM. The Pearson correlation coefficient between the two quantities is 0.003, which is consistent with no correlation (1-sided P-value = 0.48).

A previous experiment using ChIP-PCR showed that origin efficiency is strongly correlated with the number of MCMs bound at origins [114]. Those data show that, on average, there are six times more MCM on efficient origins than on inefficient origins (Supplementary Fig. 1 in [114]). A similar experiment done using ChIP-chip [113] is not consistent with the ChIP-PCR data (Fig. 4.4), possibly because of the lack of sufficient dynamic range in the ChIP-chip data.

In a collaboration to further test the MCM-n correlation, the Rhind lab recently used ChIP-seq to probe the genome-wide MCM occupancy in *S. cerevisiae* (unpublished). The resulting data are sequences that correspond to the MCM-bound DNA in a population of cells. We quantify the MCM occupancy by counting the number of sequences that fall within a radius of r_{ORI} kb around the MIM origins. To test the correlation, we ignore telomeric origins, defined as origins that are within 10 kb of the chromosomes' ends, because the fit parameters at the ends, being determined by the data from only one side (the other side has no data), are less reliable. The MCM occupancy shows significant correlation with the MIM n for a range of r_{ORI} (Fig. 4.5A). The apparent maximum in correlation at $r_{\text{ORI}} \approx$ 2.4 kb suggests the typical radius of an origin in terms of MCM spread. The

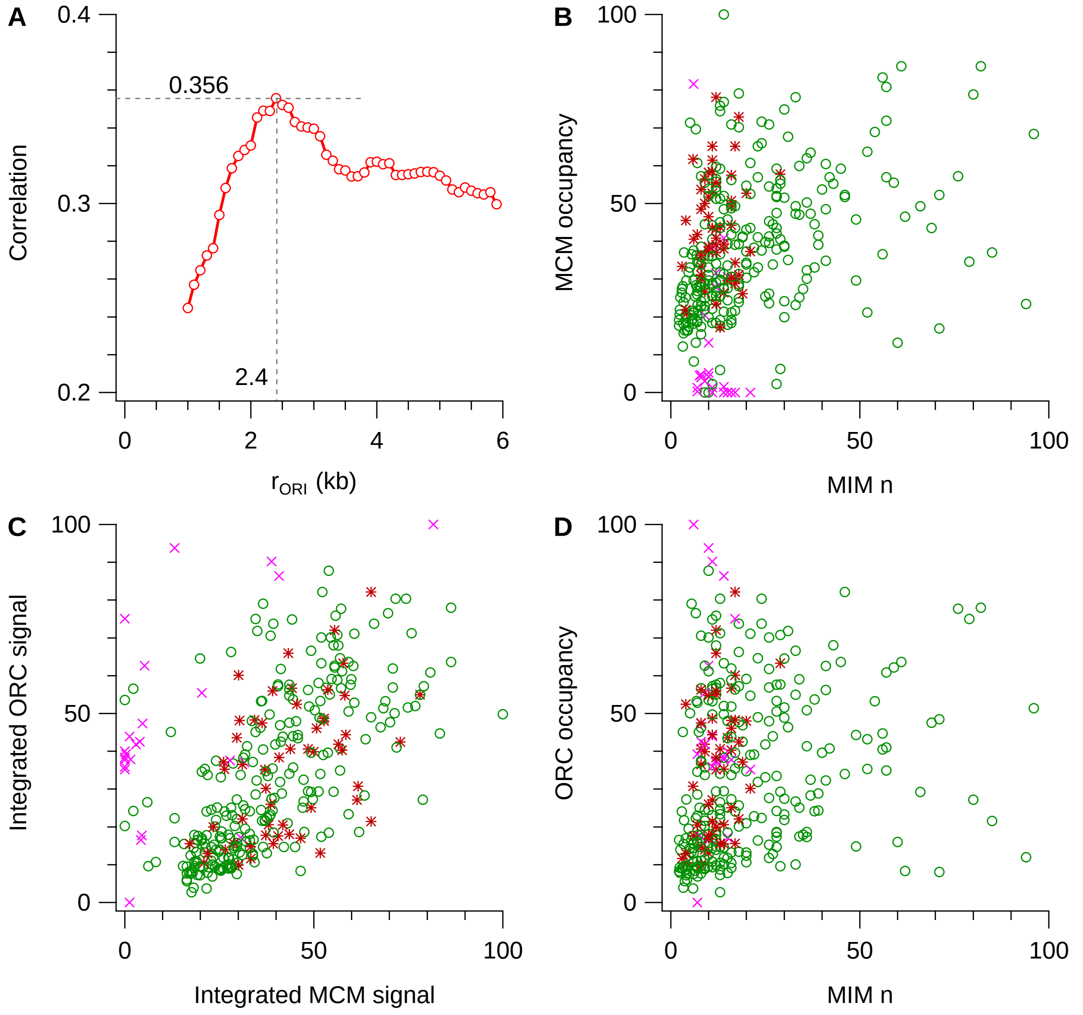


Figure 4.5: Correlations among MIM n, MCM occupancy, and ORC occupancy in budding yeast. **A**. The y-axis gives the Pearson correlation coefficient between the 313 MIM n (337 normal origins $-$ 24 telomeric origins) and MCM occupancies. The MCM occupancy of an MIM origin is calculated by counting all MCM-bound sequences that fall within a radius r_{ORI} of that origin's position. The x-axis indicates the radius r_{ORI}. Maximum correlation $=$ 0.356 at r_{ORI} =2.4 kb. **B**,**C**, and **D**. The MCM occupancy shown is calculated at r_{ORI} =2.4 kb and normalized so that the maximum occupancy is 100. The ORC signal is calculated similarly at r_{ORI}=2.4 kb and normalized to span 0–100. The MIM n are the final fit parameter values. Each plot has 337 points: 24 crosses corresponding to telomeric origins that are within 10 kb of the chromosomes' ends; 54 stars corresponding to rdp3 origins identified in [115]; and 259 circles corresponding to normal origins.

Pearson correlation between MCM occupancy and n at this r_{ORI} is 0.356 (1-sided P-value $= 2 \times 10^{-11}$; Fig. 4.5B). This significant correlation confirms the previous ChIP-PCR correlation and supports the proposal that MCM occupancy is a primary determinant of origin initiation time.

Recently, Wu and Nurse reported a correlation between the timing of initiation and the timing of ORC binding in *S. pombe* (fission yeast) that also suggests a mechanism for origin-timing control [116]. Since ORC recruits multiple MCM [99, 109], we speculate that early ORC binding provides more opportunities for MCM to be loaded, leading to the formation of early-firing origins. We found evidence that indirectly supports this speculation in *S. cerevisiae* (budding yeast). From recent ChIP-seq data [47], we calculate ORC occupancy by counting the number of ORC-bound sequences that fall within r_{ORI} =2.4 kb of the MIM origins and taking the logarithm of the total count. Assuming that sites with high ORC occupancy started loading ORC early, it is reasonable to also assume that these sites have more time to load MCM. Thus, we expect the ORC and MCM occupancy to correlate. Indeed, ORC and MCM occupancy do correlate significantly (Pearson correlation $=$ 0.638; P-value negligible; Fig. 4.5C). Interestingly, the Pearson correlation between MIM n and ORC occupancy is 0.258 (1-sided P-value $= 1 \times 10^{-6}$; Fig. 4.5D) and is close to the product of the MIM-MCM and MCM-ORC Pearson correlations ($0.356 \times 0.638 = 0.227$). This suggests that ORC indirectly controls the initiation timing through the loading of MCM and that the two steps of loading and activation of MCM are independent stochastic events.

4.3.4 Correlation with chromatin structure?

Although MCM occupancy correlates significantly with the MIM n, it is clear that there are still unexplained features. We suggested above that chromatin structure may be another contributing factor. To test this proposal, we consider the nucleosome, a compact structural feature of chromatin that consists of a segment of DNA wrapped around eight histone proteins [117]. Its stability can be modified via several processes; the analysis below focuses on acetylation and deacetylation. Acetylation of histones destabilizes the nucleosome by neutralizing the positive charge of histones [118]. The interaction between the neutralized histones and the negatively charged DNA decreases as a result, and the nucleosome disassembles. The reverse process of deacetylation recovers the positive charge of the histones, thereby promoting formation of nucleosomes. Figure 4.6 illustrates these two processes*. In higher organisms such as *H. sapiens*, histone acetylation (or, interchangeably, nucleosome acetylation) forms open chromatin structures, and these structures correlate well with gene-rich, high-transcription, earlier-replicating regions [95, 120]. The picture is that the transcription and replication machinery can more easily access these regions.

The situation is more confusing in *S. cerevisiae*. Although histone acetylation, which marks open chromatin states, correlates well with transcription activity [121], transcription activity does not correlate well with replication timing [20]. The two observations together suggest that replication timing is not affected by the "openness" of chromatin. We investigate this suggestion by testing whether the level of histone acetylation correlates with MIM n. Pokholok *et al.* used ChIP-chip to estimate the genome-wide level of nucleosome acetylation [121]. We calculated acetylation levels, which reflect chromatin accessibility, by summing up all chip intensities that lie within a radius $r_{\mathrm{ORI}} = 2.4$kb of the MIM origins. We found no correlation between the MIM n and the acetylation level (313 normal origins used; Pearson correlation $= 7 \times 10^{-4}$; 1-sided P-value = 0.49). While this result does not support our proposal that initiation timing correlates with chromatin accessibility, we note that ChIP-chip data are not reliable, perhaps because of insufficient dynamic range, as seen in Sec. 4.3.3 (compare Figs. 4.4 to 4.5B).

There is more-recent evidence based on ChIP-seq data that the "compactness" of chromatin structure affects initiation timing. Knott *et al.* used ChIP-seq to show that upon deleting rdp3, the gene for a component of histone deacetylase, over 100 non-telomeric origins initiate earlier than in wild type. Unlike the previous datasets, the data for deacetylation concern only the origins that are affected by rdp3 rather than the genome-wide deacetylation intensity [115]. Of the 106 origins affected by rdp3 levels [115], 54 coincide with an MIM origin within 5 kb. We found that these origins are relatively low in n but high in MCM occupancy (Fig. 4.5B). This matches our intuition that, although these origins consist of many initiators, the firing probability is suppressed by a more compact chromatin

*Another (perhaps more popular) view of histone acetylation is that the process causes nucleosomes to interact less with each other (instead of disassembling) to form an accessible chromatin structure. The reverse process, deacetylation, strengthens the attraction among nucleosomes to promote a compact heterochromatin structure [119]. The details of how chromatins become accessible upon histone acetylation is unimportant to our analysis here.

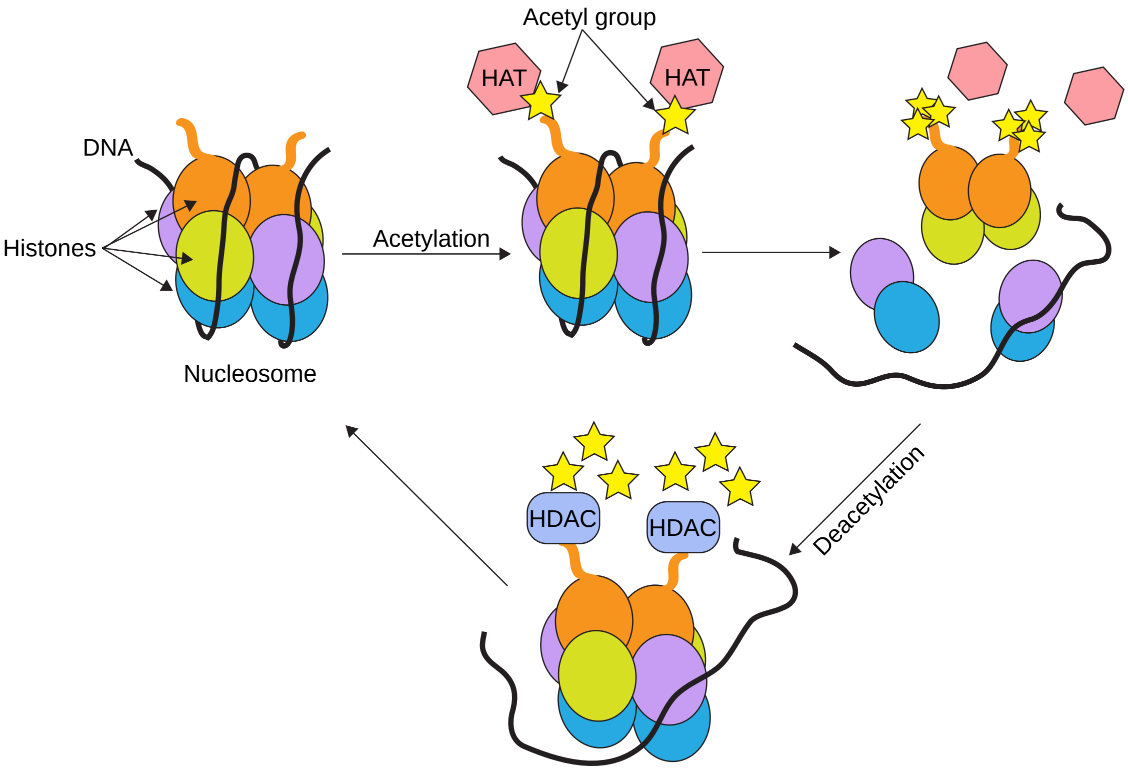


Figure 4.6: A schematic diagram of acetylation and deacetylation of nucleosomes. In the acetylation process, histone acetyltransferases (HAT) incorporates acetyl groups to two histones. The acetyl groups neutralize the histones, and the nucleosome falls apart. The DNA in this state is more accessible. In deacetylation, histone deacetylase (HDAC) removes the acetyl groups, and the nucleosome reassembles. The DNA in this state is less accessible. This figure is a modified version of Fig. 5a in [118].

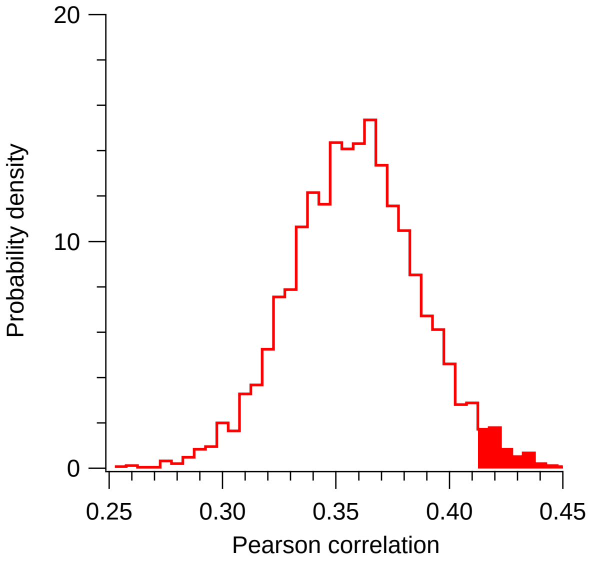


Figure 4.7: The distribution of Pearson correlation for MIM n and MCM occupancy is obtained as follows: 54 points are randomly deleted from the 313 MIM n that are non-telomeric. The corresponding 54 points are also deleted from the MCM occupancy. We repeat the random deletion 5000 times and record the resulting Pearson correlations. These correlations are used to construct the histogram (bin size = 0.005; mean = 0.356; standard deviation = 0.028). Deleting the 54 rdp3 origins resulted in a Pearson correlation of 0.408. The probability of obtaining a higher correlation by chance is equal to the shaded area (area = 0.04).

structure. Discounting the rdp3 origins, the Pearson correlation between MIM n and MCM occupancy increased from 0.356 to 0.408. Random deletion of the same number of points (54 points) leads to a distribution of Pearson correlations shown in Fig. 4.7. From this distribution, we calculated that the P-value for obtaining a correlation of 0.408 is 0.04. Thus, the effect of chromatin accessibility on initiation timing via rdp3 is significant.

In summary, from the literature and our analysis above, we can start to piece together a possible "chain of causality" for initiation timing in budding yeast:

$$\text{Sequence} \rightarrow \text{Nucleosome} \rightarrow \text{ORC} \rightarrow \text{MCM} \rightarrow \text{MIM}\, n \rightarrow t_{1/2}.$$

That is, the sequence governs the position and density of nucleosome; ORC binds to the nucleosome-free regions [47]; along with chromatin structure, the ORC loading time and efficiency determine the number of MCM loaded (Fig. 4.5C with [116]); the MCM occupancy and chromatin structure around an origin site partially explain the variation in the MIM n (Fig. 4.5B); and the MIM n describe the observed median firing times well (Fig. 4.2A). In [122], Arneodo *et al.* shows that DNA sequence encodes rich structural and dynamical information, advocating sequence as the first element of the causal chain. Although the chain is speculative and incomplete, it offers directions for future investigation. For instance, knowing that incorporating histone deacetylation improves the correlation between MCM occupancy and MIM n, one can further investigate the effects of chromatin accessibility. In particular, we propose in Chapter 7 that investigating the three-dimensional chromosome structures and the dynamic nuclear environment may reveal the missing pieces.

4.A Appendix

4.A.1 Form of firing-time distribution

Although our motivation for using the Hill equation to describe the firing-time distribution in both the SM and MIM is purely phenomenological, the Hill equation does have a physical interpretation. The equation describes cooperative binding, where the fraction of sites bound increases non-linearly as a function of ligand concentration. Comparing these quantities with those in Eq. 3.1 or 4.3, we can map concentration to time and the fraction of sites bound to the probability of initiation. The simplest interpretation of the concentration-time mapping is that the concentration of activators linearly increases with time [123]. In this scenario, the inferred MIM Hill coefficient $r^* = 3$, being greater than 1, suggests that the biochemical events leading to an initiation are cooperative. Similarly, the SM Hill coefficients $r \propto t_{1/2}^{-0.18}$ for the range of relevant $t_{1/2}$ suggests that origins fire earlier because the associated reactions are more cooperative. These physical interpretations are, of course, speculative.

Are there "more-natural" forms for the firing-time distribution? The exponential form—$\Phi(t) = 1 - \exp[-t/\tau]$, with τ being a rate constant—describing a single-step irreversible chemical reaction seems a natural candidate, as initiation is also irreversible. However, comparing the sigmoid-shaped data points in Figs. 3.1B and 3.7 with the concave form of

$\Phi(t)$, we argue that an exponential form cannot fit the time-course data well. This form is not suitable for the MIM $\Phi_o(t)$ either. Substituting it into Eq. 4.2, one sees that the effective firing-time distribution $\Phi_{\text{eff}}(t, n) = 1 - \exp[-nt/\tau]$ is still exponential and cannot generate the trend in Fig. 4.1B. In general, irreversible processes that result in purely exponential distributions, be they single-step or multiple-step, do not describe the replication kinetics in budding yeast well.

A slightly more complicated and realistic process is the Michaelis-Menten kinetics that describes the chemical reaction

$$[\text{E}] + [\text{S}] \rightleftharpoons [\text{ES}] \rightarrow [\text{E}] + [\text{P}],$$

where [E], [S], [ES], and [P] are the concentrations of the enzyme, substrate, enzyme-substrate intermediate, and product, respectively. Each arrow has an associated rate. Our preliminary study shows that, for a good range and combination of rates, the resulting distribution (the concentration of [P] as a function of time) is still roughly concave throughout time. Thus, similar to the exponential distribution, this form cannot fit the data well and is not suitable for the MIM $\Phi_o(t)$. Interestingly, Brummer *et al.* modelled the complex molecular network identified in budding yeast using rate equations and found a set of firing-time distributions whose means and standard deviations correlate [23]. In other words, although simple chemical reactions do not produce suitable distributional forms, a realistic network of reactions does. This suggests that the physics and chemistry underlying initiation may be inherently complex.

4.A.2 Parameter tables

The final fit parameter values of the SM and MIM can be found on the website: `http://www.nature.com/msb/journal/v6/n1/suppinfo/msb201061_S1.html`.

Supplementary Table I: Origin properties extracted from the genome-wide SM. For the column titles, we used the following abbreviations: "chr" for chromosome, "ori pos" for origin position, "err" for error, "pot eff" for potential efficiency, and "obs eff" for utilized efficiency[*]. Under the "Alvino," "OriDB," and "MIM" columns, 1 denotes that the origin is

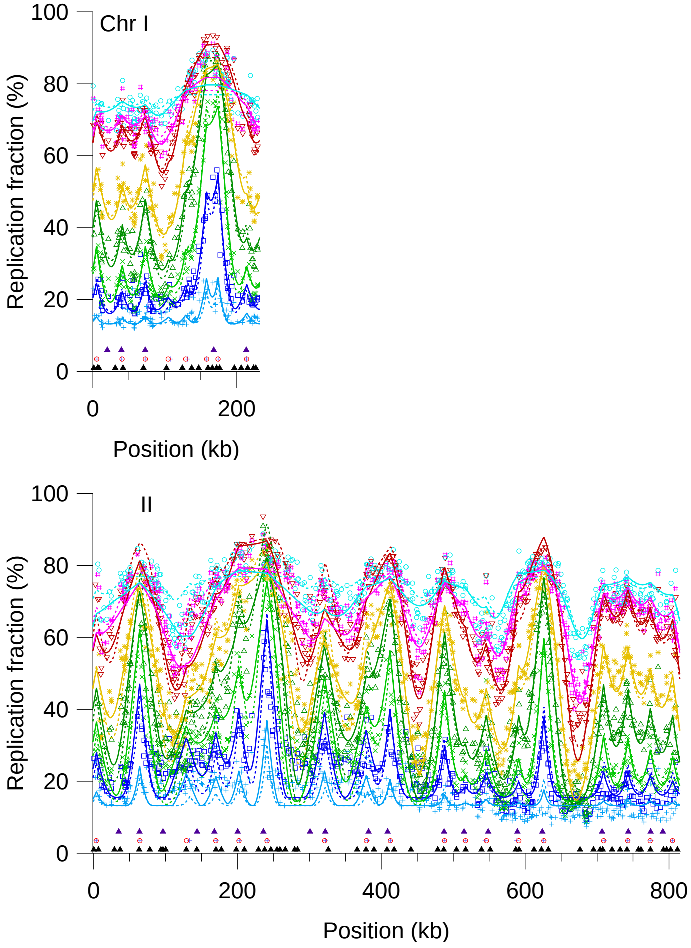
[*] When this work was published, we used "observed efficiency" instead of utilized efficiency. Later, it was pointed out that the efficiency is not observed but inferred; thus, we changed the adjective to "utilized."

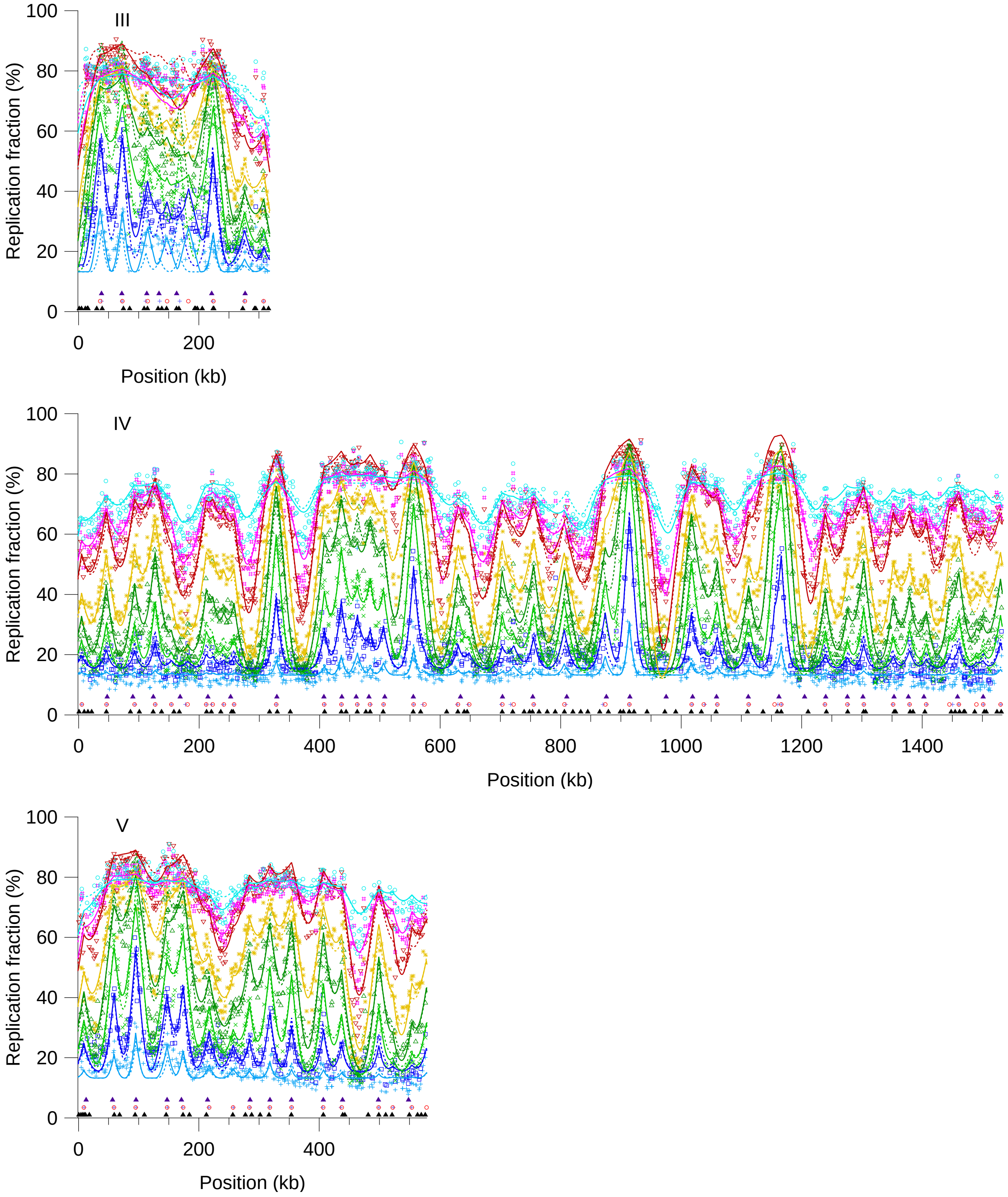
also identified in [96], in [97], and in the MIM, respectively, while 0 denotes not identified.

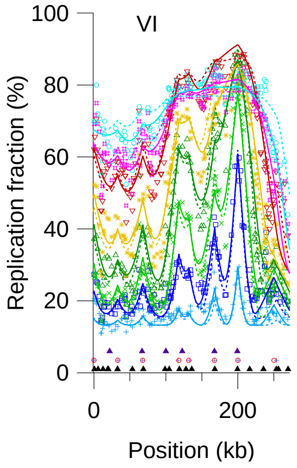
Supplementary Table II: Origin properties extracted from the genome-wide MIM. Same convention as Supplementary Table I.

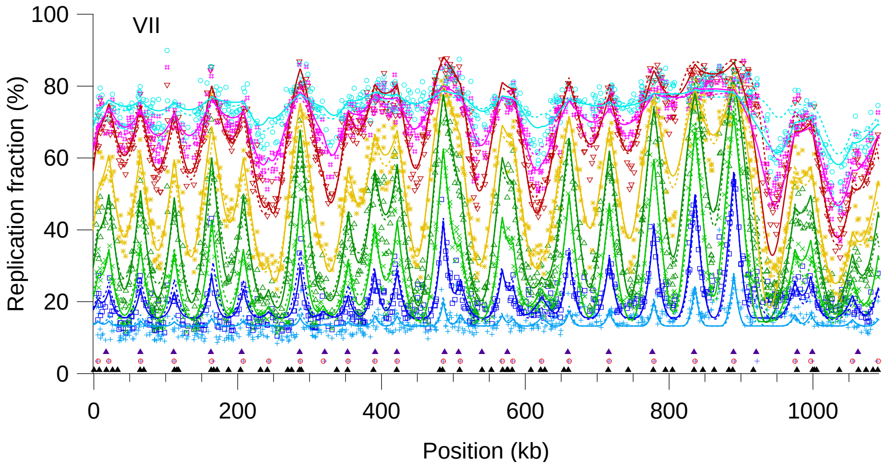
Supplementary Table III: Genome-wide parameters extracted from the SM and MIM fits. For the MIM, $t^*_{1/2}$ and r^* are used to construct the global $\Phi_o(t) = 1/[1 + (t^*_{1/2}/t)^{r^*}]$ (Eq. 4.3). The quantity $t^*_{1/2}$ plays a role that is analogous to the quantity $t_{1/2}$ for the SM model.

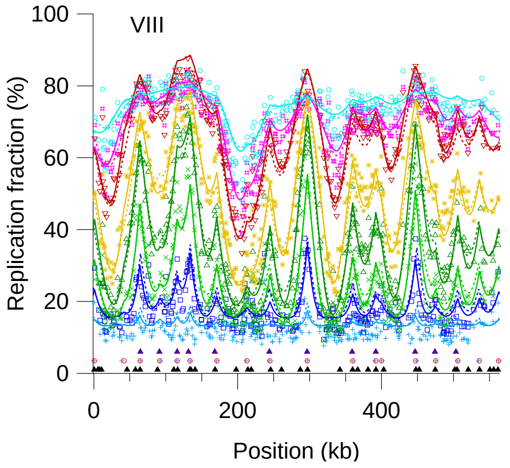
4.A.3 Whole-genome fits

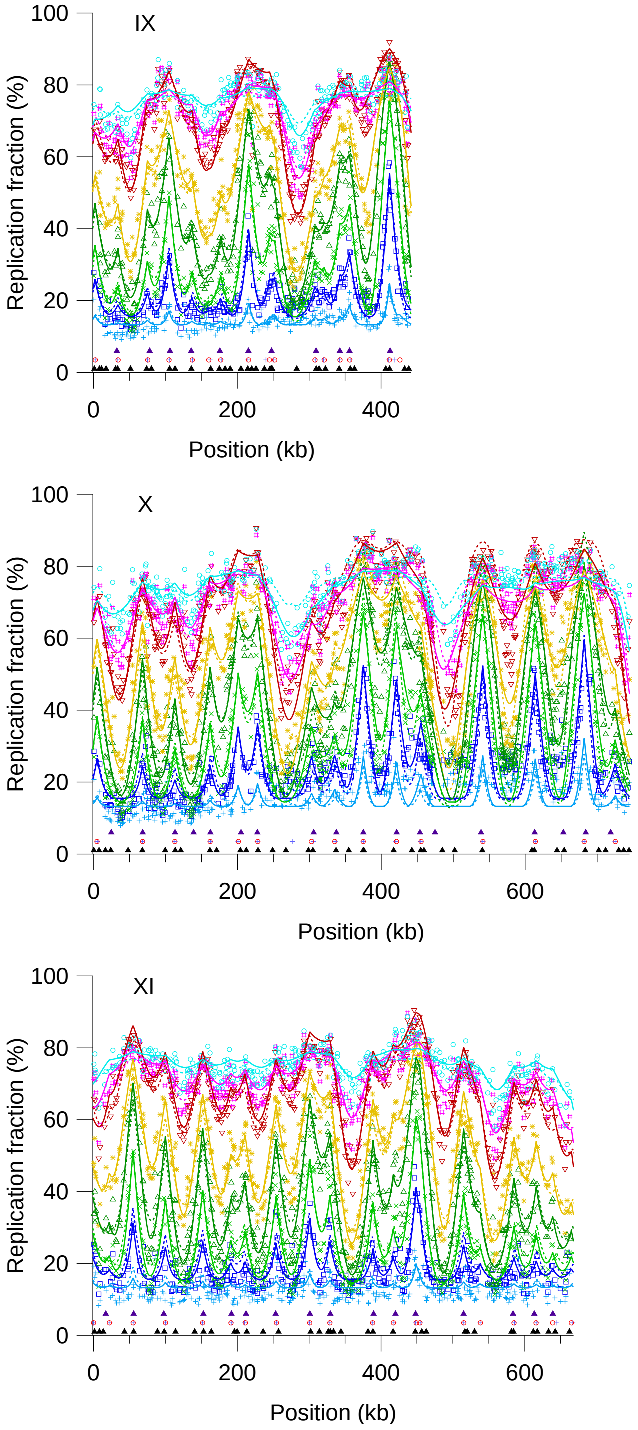


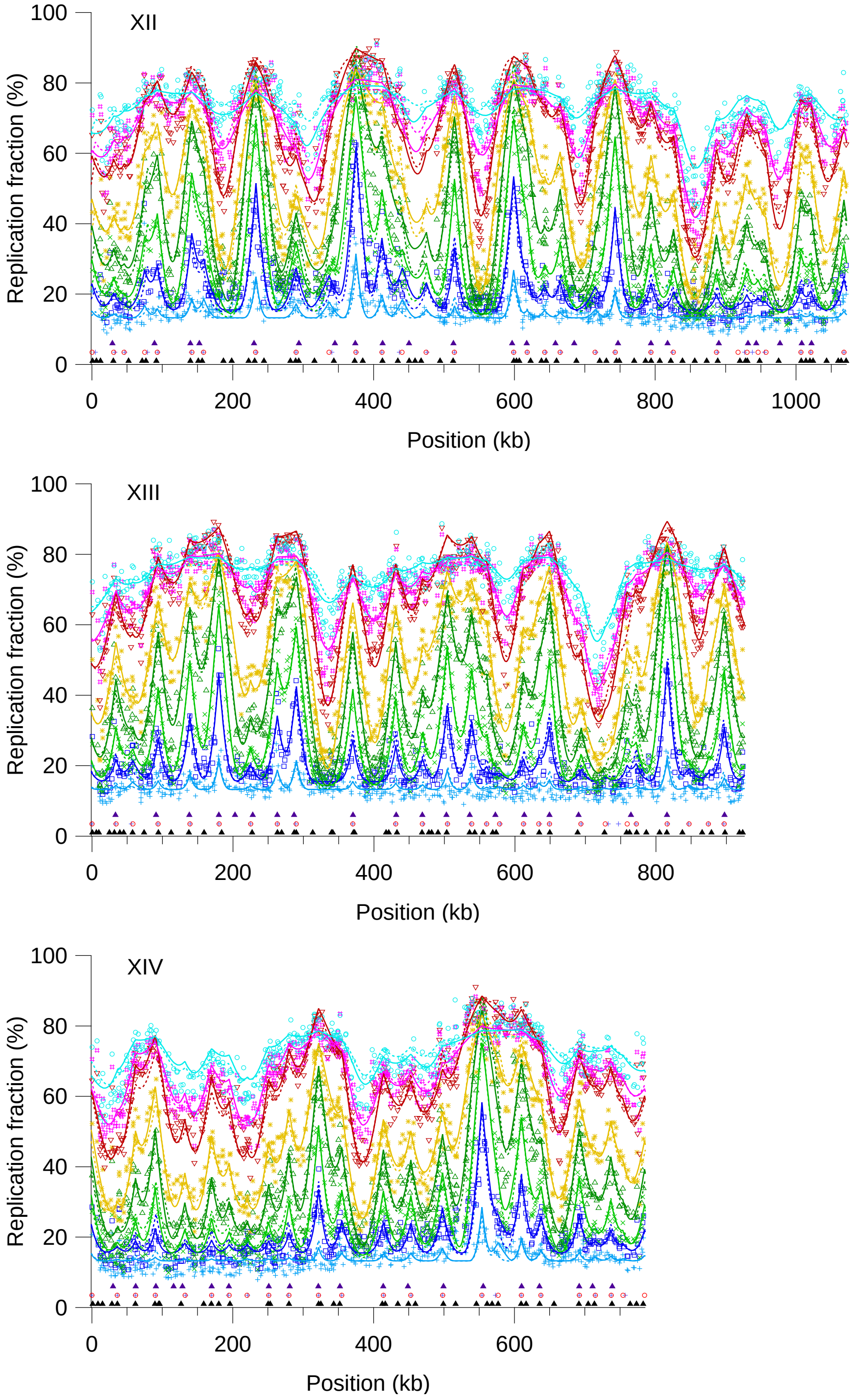

III
Replication fraction (%)
100
80
60
40
20
0
0
200
Position (kb)
IV
Replication fraction (%)
100
80
60
40
20
0
0
200
400
600
800
1000
1200
1400
Position (kb)
V
Replication fraction (%)
100
80
60
40
20
0
0
200
400
Position (kb)

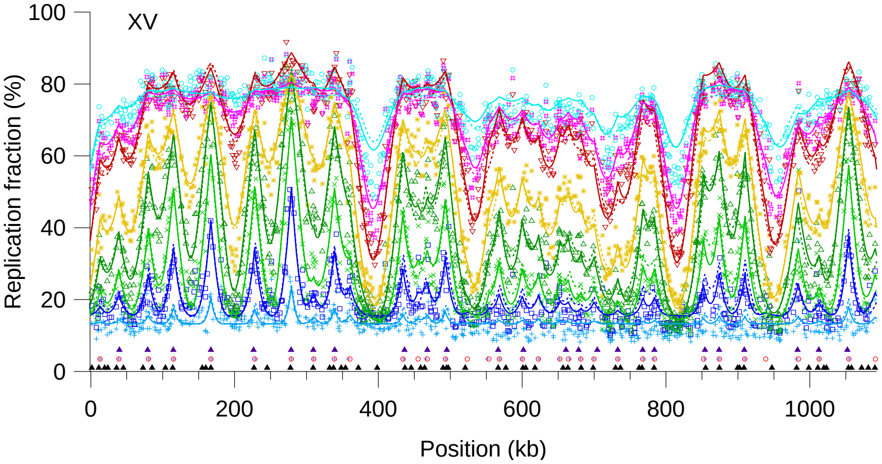

VI
Replication fraction (%)
0
20
40
60
80
100
0
200
Position (kb)


VII
Replication fraction (%)
0
20
40
60
80
100
0
200
400
600
800
1000
Position (kb)


VIII
Replication fraction (%)
0
20
40
60
80
100
0
200
400
Position (kb)


IX
Replication fraction (%)
100
80
60
40
20
0
0
200
400
Position (kb)
X
Replication fraction (%)
100
80
60
40
20
0
0
200
400
600
Position (kb)
XI
Replication fraction (%)
100
80
60
40
20
0
0
200
400
600
Position (kb)


XII
Replication fraction (%)
0
20
40
60
80
100
0
200
400
600
800
1000
Position (kb)
XIII
Replication fraction (%)
0
20
40
60
80
100
0
200
400
600
800
Position (kb)
XIV
Replication fraction (%)
0
20
40
60
80
100
0
200
400
600
Position (kb)



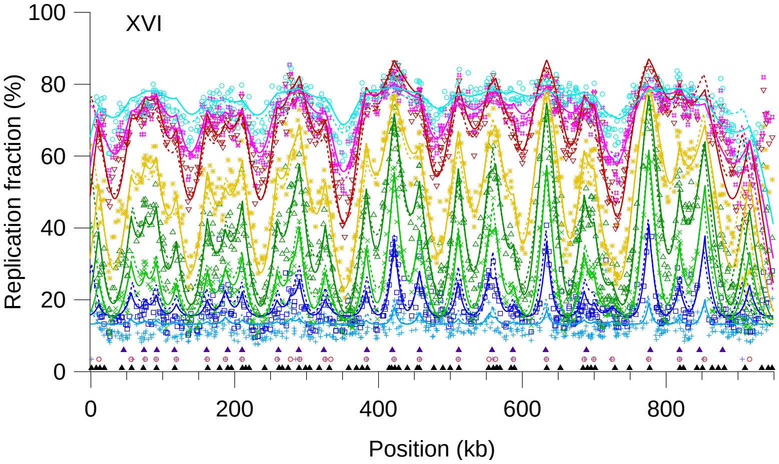


Figure 4.8: Genome-wide SM and MIM fits, separately shown for each chromosome. Roman numbers correspond to chromosome number. The x-axis denotes the position along the chromosome. At the bottom, upper row of solid triangles denote origin positions identified in [96]; middle row of open circles denote the estimated origin positions in the sigmoid model (SM); and the lower row of triangles correspond to origins in the OriDB database [97]. Markers are data; solid lines are fits from SM; dotted lines are fits from MIM. The eight curves from bottom to top correspond to the replication fraction $f(x)$ at 10, 15, 20, 25, 30, 35, 40, and 45 min after release from the *cdc7-1* block.

Chapter 5

Reconstructing the Replication Program from Asynchronous Replication Experiments

In Chapters 3 and 4, we extracted the spatiotemporal replication program for budding yeast from a time-course microarray experiment. Even though the time-course microarray technique is arguably the most informative among the ones mentioned in Sec. 1.3, it is expensive, time-consuming, and requires the ability to synchronize the cell culture probed. In this chapter, we focus on the more-accessible FACS-microarray technique. We recall that fluorescence-activated cell sorting (FACS) is a technique for selecting cells that are progressing through S phase in an asynchronous cell population. The major difference between the two techniques is that the former provides information on the replication fraction $f(x, t)$ as a function of genome position and time, whereas the latter provides information on the spatially averaged replication fraction profile $f(t)$ and the temporally averaged replication fraction profile $f(x)$. We investigate methods to extract the spatiotemporal replication program from simulated FACS-microarray data. That is, from measurements of $f(t)$ and $f(x)$, we seek to estimate $f(x, t)$ and the initiation rate $I(x, t)$. Our hope is that these proof-of-principle simulation studies can eventually be extended to real FACS-microarray experiments.

5.1 Introduction

In Sec. 1.3, we introduced several experimental techniques. Only two of the techniques, time-course microarray and FACS-microarray, are both genome wide and mappable. The time-course microarray, being synchronized, is the more informative technique and allows direct application of Eq. 2.5 to the data. It is with such high-quality data that we were able to extract detailed features of the spatiotemporal replication program in budding yeast in Chapters 3 and 4. Even though time-course microarray is powerful, it has limited applicability because synchronization of cell cycles in a population is generally very difficult for eukaryotes other than yeast. Even when possible, synchronization methods often involve genetic mutations that may alter the replication program in unintended ways.

For these organisms, FACS-microarray is the method of choice. As the name suggests, the technique involves two steps. The first is to sort the cells in a population into different categories based on their DNA content via FACS (Fig. 5.1A). There are many variants of the technique. Here, we focus on the simplest, which sorts the cells into only two categories: replicating or non-replicating. Replicating cells have DNA content between one copy (1C) and two copies (2C) of DNA; non-replicating cells have either 1C or 2C of DNA. One can then sort "replicating" cells by selecting those whose intensities of labelled DNA lie between a lower and upper threshold (Fig. 1.6B). The replicating cells are then hybridized onto a microarray (or, more likely, sequenced in 2012) to generate a spatial replication profile. In terms of the replication fraction $f(x,t)$ defined in the beginning of Sec. 1.3, the resulting spatial replication profile $f(x)$ is equal to $f(x,t)$ averaged over time. The FACS data, as we will discuss in Sec. 5.2 below, are intimately connected with $f(t)$, which is defined to be $f(x,t)$ averaged over genome position[*]. In analogy to the technique of tomography [124], the FACS microarray produces two perpendicular projections of the time-course microarray $f(x,t)$—the x and t projections.

Similar to the case of time-course microarray, our goal here is to extract the replication program $I(x,t)$ from FACS-microarray data. We assume, as in previous chapters, that the fork velocity v is constant. In parts of the analysis, we shall also assume that the constant velocity is known in order to further simplify the treatment. All of the methods that we

[*] Here, we use a loose notation where $f(x)$, $f(t)$, and $f(x,t)$ denote three different functions. We use the different arguments to distinguish among them.

present are tested on simulated data.

The chapter is organized as follows: In Sec. 5.2, we present an analysis of FACS data. The goal is to first obtain the replication fraction $f(t)$ and then the initiation rate $I(t)$ as a function of time. In Sec. 5.3, we analyze the microarray profile $f(x)$ by using a time-averaged version of Eq. 2.5. The method is essentially the same as the fit presented in Chapter 3. In Sec. 5.4, we present a non-parametric inversion method that takes $f(t)$ and $f(x)$ as inputs and generates the spatiotemporal $I(x,t)$ as the output.

5.2 Analysis of FACS data

FACS is a very popular and efficient technique for determining the relative DNA content of each cell in a cell culture. Present-day flow cytometers can process $\mathcal{O}(10^4)$ cells per second [59]. The basic experimental procedures involve staining the cells' DNA with fluorophores, flowing the cells one by one through a narrow channel, shining a laser beam at the cell, detecting the fluorescence intensity, and directing via fluid flow the cell to a particular storage chamber based on the intensity signal (Fig. 5.1A). The intensity of the signal reflects the amount of DNA in the detected cell, as a cell with more DNA will incorporate more fluorophores. Given an asynchronous cell culture that is growing under normal conditions, a typical FACS-experiment result is a histogram of DNA content, as illustrated in Fig. 5.1B.

The standard, qualitative interpretation splits the histogram in Fig. 5.1B into three parts: The left peak corresponds to cells that have 1C DNA and have not entered into S phase (G1 phase). The right peak corresponds to cells that have 2C DNA and have finished S phase (G2 and M phase). The wide valley in between then corresponds to cells that have DNA content between 1C–2C and are replicating (S phase). The area under each part reflects the duration of the corresponding phase of the cell cycle. In the discussion below, we will work with discretized quantities because the data are always discretized. We define the experimental FACS histogram $\mathbf{h^e}_f$ to be the fraction of cells having DNA content between $[1+f, 1+f+\Delta f)$. Note that f is the replication fraction, and the DNA content is $1+f$.

In an ideal experiment where the DNA content of each cell can be measured precisely, the peaks at 1C and 2C would each occupy only one bin. We denote ideal FACS histogram

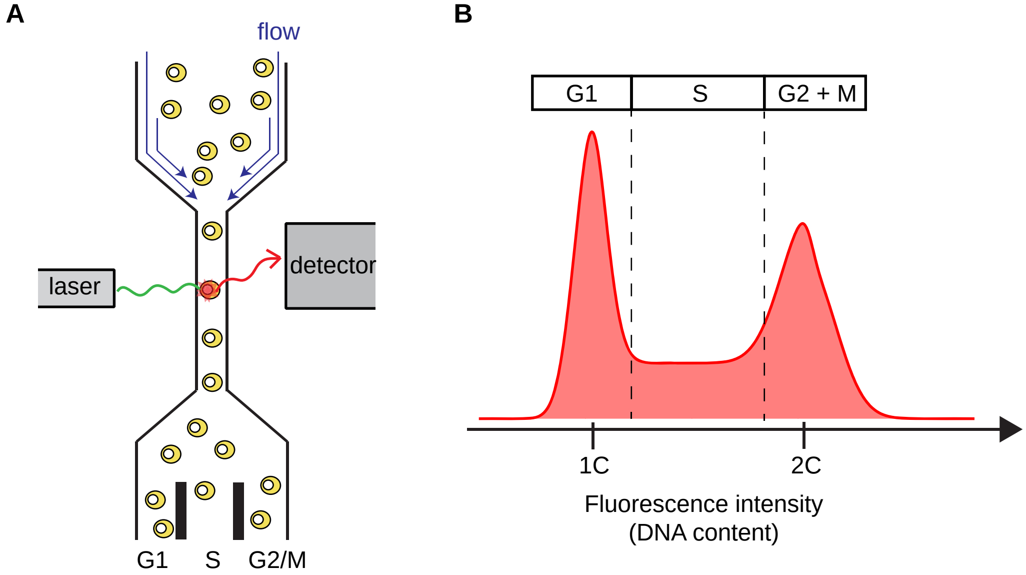


Figure 5.1: A schematic diagram of fluorescence-activated cell sorting (FACS) taken from Fig. 1.4. **A**. The setup. **B**. The resulting histogram.

as $\mathbf{h^i}_f$. The quantity $\mathbf{h^i}_f$ can be related to the replication fraction $f(t)$ via

$$\mathbf{h^i}_f = \frac{1}{T_{cc}} \left[t(f + \Delta f) - t(f) \right] , \tag{5.1}$$

where $t(f)$ is the inverse of $f(t)$, and T_{cc} is the total cell cycle time. We note that the first bin (which contains $f = 0$) and the last bin (which contains $f = 1$) are not well defined by Eq. 5.1 because the existence of finite G1, G2, and M phases is not built into the definition of $f(t)$. Therefore, these two bins in the ideal FACS histogram need to be generated using other information. We also note that one can rewrite Eq. 5.1 with a scaled time, without the factor of T_{cc}. We retain it in the definition in order to make a clearer connection with experiment. The meaning of the FACS histogram is more apparent when Eq. 5.1 is rewritten in the limit $\Delta f \to 0$:

$$\mathbf{h^i}(f) = \left(\frac{df}{dt} \right)^{-1} . \tag{5.2}$$

Equation 5.2 shows that cells that progress slowly through a given part of S phase spend more time at roughly the same f; thus, more cells are sampled in the parts of S phase where $f(t)$ is changing most slowly—namely, at the beginning and at the end of S phase. Underlying Eq. 5.2 is also the assumption that the cells are uniformly sampled in the cell cycle time.

The broad peaks at 1C and 2C seen in Fig. 5.1B result from a combination of factors, including fluctuations in the number of fluorophores incorporated, fluctuations in the number of photons emitted by the fluorophores, and amplification of noise in the detector. We describe the combined effect of these noise sources by a point-spread function. The function describes how the measured DNA content is distributed around the true content. It has been proposed that a satisfactory form for the point-spread function is $N(f, \gamma f)$, where $N(\mu, \sigma)$ is a normal distribution with mean μ and standard deviation σ, and γ is a constant [125, 126]. It is not surprising that the standard deviation increases with replication fraction because many of the processes that give rise to the spread, such as the incorporation of fluorophores, are Poisson processes whose standard deviations scale with the amount of DNA. In the analysis below, we used the slightly different form $N(f, \gamma f + c)$, where c is another constant. This choice reflects our observation that the spread at 2C is usually smaller than twice the spread at 1C. In effect, we assume the existence of another uncorrelated noise source that is independent of the DNA content (e.g., non-specific binding of the fluorophores).

In a discretized version, the point-spread function becomes a point-spread matrix $\mathbf{M}$, whose row i elements are formed from $N(i, \gamma i + c)$:

$$\mathbf{M}_{ij} = \frac{e^{-\frac{1}{2}\left(\frac{j-i}{\gamma i + c}\right)^2}}{(\gamma i + c)\sqrt{2\pi}}. \tag{5.3}$$

The two FACS histograms are then related via matrix multiplication,

$$\mathbf{h^e} = \mathbf{M}\mathbf{h^i}. \tag{5.4}$$

Note that $\mathbf{h^i}$ has n_{id} elements whose value f lies strictly in the range $[0, 1]$. On the other hand, $\mathbf{h^e}$, being "corrupted" by the point-spread function, can cover a larger range of values. (Typically, they are contained in $[-1, 2]$.) Because the histogram bin size is the same in $\mathbf{h^i}$ and $\mathbf{h^e}$, the number of elements of $\mathbf{h^e}$, n_{ex}, is greater than n_{id}. The point-spread matrix $\mathbf{M}$ is then a n_{ex}-by-n_{id} matrix. Throughout the chapter, bold capital letters are used for matrices, and bold small case letters for vectors.

Having established the relationship between $f(t)$ and $\mathbf{h^e}$ via Eqs. 5.1 and 5.4, we consider possible ways to extract the initiation rate and fork velocity from FACS data. Since FACS data provides no spatial information, one can at most extract a time-dependent rate $I(t)$ and velocity $v(t)$ from the data. One way to do this is to construct a model that parameterizes the point-spread function, the heights of the first and last bin, the initiation rate $I(t)$, and the velocity $v(t)$. Then, one fits this model to $\mathbf{h^e}$ to extract the parameters. Another way is to perform a two-step inversion: First, one deconvolves the experimental FACS histogram into an ideal FACS histogram. Then, one transforms the ideal FACS histogram into $I(t)$ and $v(t)$. Below, we present a simulation study and use the latter strategy, where we deconvolve $\mathbf{h^e}$ into $\hat{\mathbf{h}}^{\mathbf{i}}$ and then invert a discretized $I(t)$ from $\hat{\mathbf{h}}^{\mathbf{i}}$. The hat denotes an estimate throughout the chapter.

5.2.1 Deconvolving FACS histogram

We first simulate a FACS experiment. The simulation involves:

1. Assigning each cell in the cell culture a time in the cell cycle. The time is generated from a uniform distribution. The cell culture contains 10^6 cells.

2. Determining the cell's DNA content. We let the G1, S, and G2+M phases occupy 21, 28, and 51% of the cell cycle time, respectively. If the assigned time for a cell is in G1 phase, the cell is assigned a 1C DNA content. If in G2 or M, the cell is assigned a 2C DNA content. If the cell is in S phase, the DNA content is $1 + f$, where f is the replication fraction simulated using the routines mentioned in Sec. 2.3. The length of the genome for each cell is 10^3 kb. The initiation rate $I(t)$ linearly increases in time. The velocity v is 1 kb/min.

3. Making a histogram of the DNA content. This produces the ideal FACS histogram.

4. Smoothing the ideal histogram with the point-spread function. The point-spread function varies between $N(0, 0.08)$ at the G1 peak and $N(1, 0.1)$ at the G2+M peak. Specifically, we choose $\gamma = 0.02$ and $c = 0.08$. The experimental FACS histogram is obtained via Eq. 5.4.

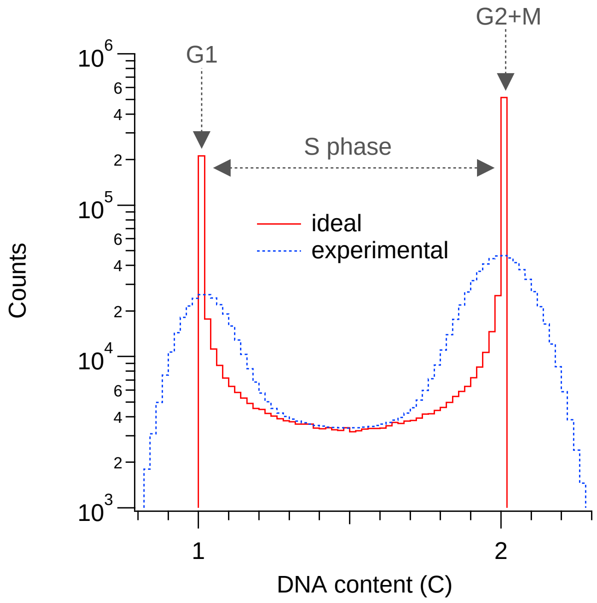


Figure 5.2: Simulated FACS histograms. The ideal FACS histogram $\mathbf{h}^{\mathbf{i}}$ is plotted with dotted line, and the experimental FACS histogram $\mathbf{h}^{\mathbf{e}}$ is plotted with solid line. Bin size $= 0.02$ C.

The resulting FACS histograms are shown in Fig. 5.2.

To set up the analysis, we want to solve for $\mathbf{h}^{\mathbf{i}}$ in Eq. 5.4, where $\mathbf{h}^{\mathbf{i}}$ is an n_{id}-element column vector, $\mathbf{h}^{\mathbf{e}}$ is an n_{ex}-element column vector, M is an n_{ex}-by-n_{id} matrix, and $n_{\mathrm{ex}} > n_{\mathrm{id}}$. The point-spread matrix M is not directly invertible because it is not square. The standard procedure is to solve the equation in the least-squares sense, by finding the $\hat{\mathbf{h}}^{\mathbf{i}}$ that minimizes $(\mathrm{M}\hat{\mathbf{h}}^{\mathbf{i}} - \mathbf{h}^{\mathbf{e}})^{\mathrm{T}}(\mathrm{M}\hat{\mathbf{h}}^{\mathbf{i}} - \mathbf{h}^{\mathbf{e}})$, where $^{\mathrm{T}}$ denotes the transpose. The solution is

$$\hat{\mathbf{h}}^{\mathbf{i}} = (\mathrm{M}^{\mathrm{T}}\mathrm{M})^{-1}\mathrm{M}^{\mathrm{T}}\mathbf{h}^{\mathbf{e}}, \tag{5.5}$$

where $^{-1}$ denotes the inverse. For simplicity, we assume that the point-spread matrix M is given. In real applications, one has to estimate M either from $\mathbf{h}^{\mathbf{e}}$ or from independent measurements. Figure 5.3A shows the solution obtained via Eq. 5.5.

The large fluctuations shown in Fig. 5.3A are clearly undesirable. Such fluctuations are inherent to inverting the smoothing matrix M. Following Höcker and Kartvelishvili [127], we clarify the underlying issue by investigating the singular value decomposition (SVD)

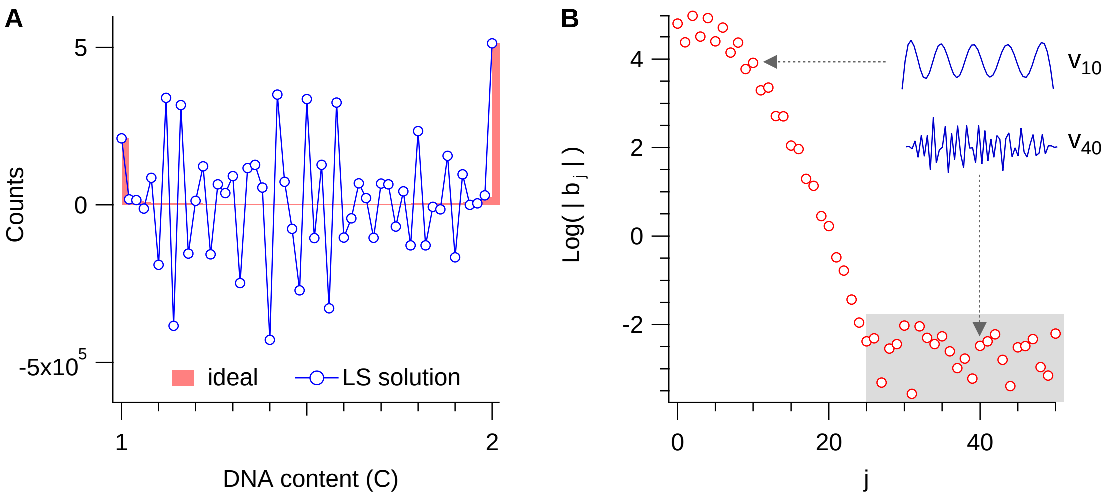


Figure 5.3: Deconvolution of FACS histograms by a naive least-squares method. **A**. The ideal FACS histogram $\mathbf{h}^\mathbf{i}$ in Fig. 5.2 is plotted on linear scale here. The deconvolved histogram obtained via Eq. 5.5 is plotted with connected markers. **B**. The y-axis is the $\log$ of the absolute value of b_j defined under Eq. 5.6. The x-axis is j and ranges from $[1, n_{\mathrm{id}}]$. At $j = 25$, $\log(|b_j|)$ changes from being signal dominated to being noise dominated (shaded area). The inset shows the singular vectors $\mathbf{v}_{10}$ and $\mathbf{v}_{40}$, also defined under Eq. 5.6.

of $\mathbf{M}$. Briefly, the SVD generalizes the notion of eigenvalue decomposition to a matrix, with $\mathbf{M} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{U}$ is n_{ex}-by-n_{id} and column orthonormal, $\mathbf{S}$ is n_{id}-by-n_{id} and diagonal, and $\mathbf{V}$ is n_{id}-by-n_{id} and orthonormal. The values along the diagonal of $\mathbf{S}$ are called singular values and are ranked from largest to smallest going from top left to bottom right.

Using SVD, we rewrite Eq. 5.5 as

$$\hat{\mathbf{h}}^\mathbf{i} = \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{h}^\mathbf{e} = \mathbf{V}\mathbf{S}^{-1}\mathbf{b} = \sum_{j=1}^{n_{\mathrm{id}}} \frac{b_j}{s_j}\mathbf{v}_j, \tag{5.6}$$

where b_j is the j^{th} element of $\mathbf{b}$, s_j is the j^{th} singular value of $\mathbf{M}$, and $\mathbf{v}_j$ is the j^{th} column of $\mathbf{V}$. Equation 5.6 shows that the deconvolved $\hat{\mathbf{h}}^\mathbf{i}$ is formed by bases $\mathbf{v}_j$ with coefficients b_j/s_j. A general feature of smoothing kernels, such as $\mathbf{M}$, is that their singular values span many orders of magnitude; hence, the inversion problem involving $\mathbf{M}$ is ill conditioned. Another feature of smoothing kernels is that the singular vectors—the columns of $\mathbf{U}$ and

$\mathbf{V}$—associated with small singular values are in general quickly oscillating (Figure 5.3B inset).

What do these properties imply? The b_j in $\mathbf{b} = \mathbf{U}^{\mathrm{T}}\mathbf{h}^{\mathrm{e}}$ are the projection coefficients of $\mathbf{h}^{\mathrm{e}}$ onto the columns of $\mathbf{U}$. For reasonably smooth $\mathbf{h}^{\mathrm{e}}$, one expects the magnitude of b_j to decrease with j because the frequency of oscillation in the columns of $\mathbf{U}$ increases with j. As j increases, the coefficient b_j should eventually be dominated by contributions from the experimental noise in $\mathbf{h}^{\mathrm{e}}$. (Here, the noise in $\mathbf{h}^{\mathrm{e}}$ is due to the finite number of cells simulated.) Thus, $\log(|b_j|)$ (or $|b_j|$) vs. j should decrease for $j \leq k$ and fluctuate about a constant for $j > k$, where the index k separates the signal-dominated values from the noise-dominated values (Fig. 5.3B) [127]. In terms of basis representation, the "change in behaviour" implies that the singular vectors $\mathbf{v}_j$ for $j > k$ are statistically insignificant bases for $\mathbf{h}^{\mathbf{i}}$ and should have small weights in Eq. 5.6. This is, however, not the case because the $1/s_j$ factor increases the weight of $\mathbf{v}_j$ drastically as $j \to n_{\mathrm{id}}$. The least-squares solution is thus dominated by the insignificant, rapidly oscillating $\mathbf{v}_j$, and the result is the wildly fluctuating solution seen in Fig. 5.3A.

As recommended by Höcker and Kartvelishvili [127], we add a regularization term to the simple least-squares equation:

$$\text{Minimize } J = (\mathbf{M}\hat{\mathbf{h}}^{\mathbf{i}} - \mathbf{h}^{\mathrm{e}})^{\mathrm{T}}(\mathbf{M}\hat{\mathbf{h}}^{\mathbf{i}} - \mathbf{h}^{\mathrm{e}}) + \tau(\mathbf{C}\hat{\mathbf{h}}^{\mathbf{i}})^{\mathrm{T}}(\mathbf{C}\hat{\mathbf{h}}^{\mathbf{i}}), \tag{5.7}$$

where $\mathbf{C}$ is the regularizer matrix, and τ determines the relative weight of the regularization. A common choice for $\mathbf{C}$ is the numerical second derivative, given by

$$\mathbf{C} = \begin{bmatrix} -1 & 1 & 0 & \cdots & & & \\ 1 & -2 & 1 & 0 & \cdots & & \\ 0 & 1 & -2 & 1 & 0 & \cdots & \\ & \ddots & \ddots & \ddots & \ddots & & \\ & \cdots & 0 & 1 & -2 & 1 & \\ & & \cdots & 0 & 1 & -1 & \end{bmatrix}. \tag{5.8}$$

With this choice, minimizing the second term in Eq. 5.7 is equivalent to minimizing the total bin-to-bin curvature of $\hat{\mathbf{h}}^{\mathbf{i}}$. For $\tau \to 0$, the solution to Eq. 5.7 approaches that of Eq. 5.5 and is wildly fluctuating. For $\tau \to \infty$, the regularization term dominates, and the solution is a flat histogram with zero bin-to-bin curvature everywhere. For finite τ, the larger-scale, overall trend of $\hat{\mathbf{h}}^{\mathbf{i}}$ is determined by the first term in Eq. 5.7, while the smaller-scale, bin-to-bin features are set by the second term in Eq. 5.7. In particular, the regularization term in Eq. 5.7 reflects our assumption that $\hat{\mathbf{h}}^{\mathbf{i}}$ should be smooth on the bin-to-bin scale. Fluctuations in $\hat{\mathbf{h}}^{\mathbf{i}}$ on this scale imply rapid temporal changes in the initiation rate, which have not been observed in any organisms.

To find the solution of Eq. 5.7 in the form of Eq. 5.6, we define $\tilde{\mathbf{M}} \equiv \mathbf{M}\mathbf{C}^{-1}$ and $\tilde{\mathbf{h}}^{\mathbf{i}} \equiv \mathbf{C}\hat{\mathbf{h}}^{\mathbf{i}}$ and rewrite Eq. 5.7 as

$$\text{Minimize} \;\; J = (\tilde{\mathbf{M}}\tilde{\mathbf{h}}^{\mathbf{i}} - \mathbf{h}^{\mathbf{e}})^{\mathrm{T}}(\tilde{\mathbf{M}}\tilde{\mathbf{h}}^{\mathbf{i}} - \mathbf{h}^{\mathbf{e}}) + \tau(\tilde{\mathbf{h}}^{\mathbf{i}})^{\mathrm{T}}(\tilde{\mathbf{h}}^{\mathbf{i}}). \tag{5.9}$$

By setting $\partial J/\partial\tilde{\mathbf{h}}^{\mathbf{i}} = 0$, we obtain

$$\tilde{\mathbf{h}}^{\mathbf{i}} = (\tilde{\mathbf{M}}^{\mathrm{T}}\tilde{\mathbf{M}} + \tau\mathbf{I})^{-1}\tilde{\mathbf{M}}^{\mathrm{T}}\mathbf{h}^{\mathbf{e}}, \tag{5.10}$$

where $\mathbf{I}$ is the identity matrix. Defining the SVD: $\tilde{\mathbf{M}} = \tilde{\mathbf{U}}\tilde{\mathbf{S}}\tilde{\mathbf{V}}^{\mathrm{T}}$, we find

$$\hat{\mathbf{h}}^{\mathbf{i}} = \mathbf{C}^{-1}\tilde{\mathbf{V}}(\tilde{\mathbf{S}}^2 + \tau\mathbf{I})^{-1}\tilde{\mathbf{S}}\tilde{\mathbf{U}}^{\mathrm{T}}\mathbf{h}^{\mathbf{e}} = \sum_{j=1}^{n_{\mathrm{id}}} \frac{\tilde{b}_j\tilde{s}_j}{\tilde{s}_j^2 + \tau}[\mathbf{C}^{-1}\tilde{\mathbf{V}}]_j, \tag{5.11}$$

where $\tilde{b}_j$ is the j^{th} element of $\tilde{\mathbf{U}}^{\mathrm{T}}\mathbf{h}^{\mathbf{e}}$, $\tilde{s}_j$ is the j^{th} singular value of $\tilde{\mathbf{M}}$, and $[\mathbf{C}^{-1}\tilde{\mathbf{V}}]_j$ denotes the j^{th} column of $\mathbf{C}^{-1}\tilde{\mathbf{V}}$. We note that the solution now relates to $\tilde{s}_j/(\tilde{s}_j^2 + \tau)$ instead of simply $1/s_j$. The factor τ acts as a threshold to eliminate the contributions from singular vectors corresponding to $\tilde{s}_j \ll \tau$.

Although we assume that we know $\mathbf{M}$, $\mathbf{C}$, and $\mathbf{h}^{\mathbf{e}}$, we still need to choose τ in order to use Eq. 5.11. Following the suggestion in [127], a reasonable choice for τ is $\tilde{s}_k^2$, where k is the index at which noise begins to dominate (Fig. 5.3B). This choice of τ effectively suppresses the contribution from the insignificant, quickly oscillating singular vectors $\mathbf{v}_j$ associated with $j > k$. The plot of $\log(|\tilde{b}_j|)$ vs. j is very similar to Fig. 5.3B and reveals that the transition occurs at $k = 23$. Using $\tau = \tilde{s}_{k=23}^2$, we obtain Fig. 5.4A.

Figure 5.4A and B show that the deconvolved $\hat{\mathbf{h}}^{\mathbf{i}}$ matches very well the input $\mathbf{h}^{\mathbf{i}}$. We note that $\hat{\mathbf{h}}^{\mathbf{i}}$ has small bin-to-bin fluctuations. Although these fluctuations are small compared to the overall structure of the histogram, they would be amplified if one were to invert

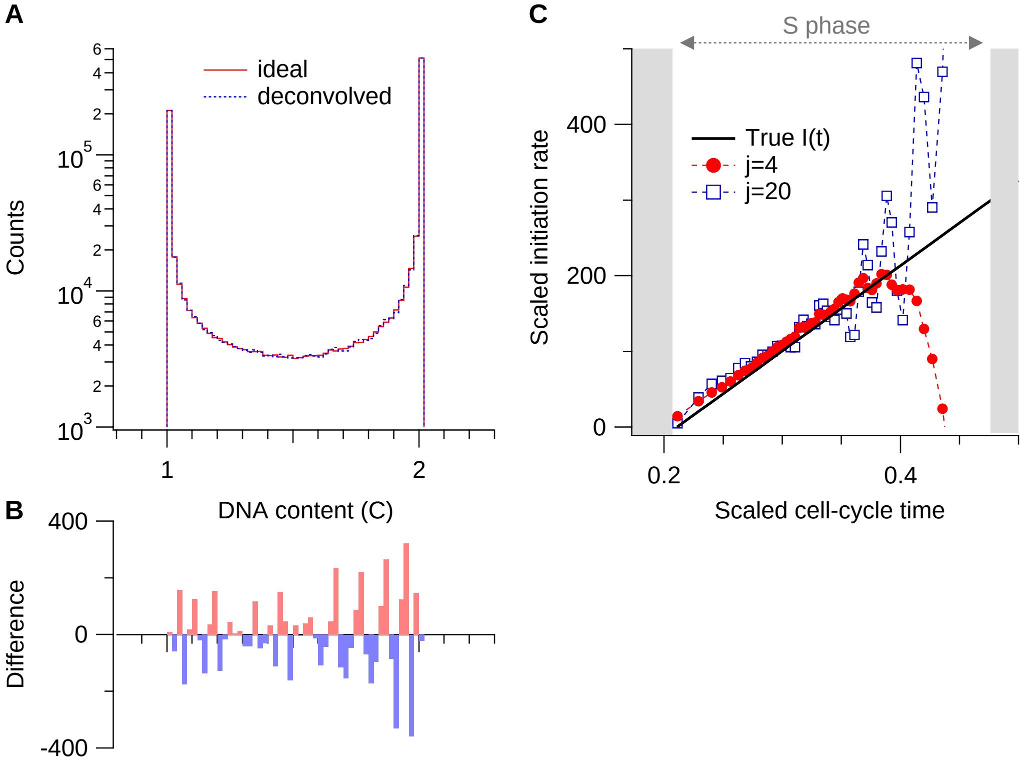


Figure 5.4: Deconvolution of FACS histograms with regularization. **A**. The ideal and deconvolved FACS histograms. **B**. The difference between the deconvolved and ideal histograms from A. The x-axis is same as A. **C**. Direct inversion of the initiation rate $I(t)$ from the deconvolved FACS histogram with regularization. Solid line represents the true $I(t)$ used in the simulation. Here, j denotes that the j^{th} singular value is used as the threshold; i.e., $\tau = s_j^2$. The x-axis is scaled so that the length of cell cycle is 1. The initiation rate $I(t)$ is scaled accordingly.

$\hat{\mathbf{h}}^{\mathbf{i}}$ to obtain the initiation rate $I(t)$. A consideration of Eqs. 2.15 and 5.1 shows that the inversion involves taking two time derivatives of $-\ln[1-f(t)]$ and that the time derivative depends on $\hat{\mathbf{h}}^{\mathbf{i}}$. Intuitively, since the noise amplification comes from the numerical derivatives and since numerical derivatives are linear operations, one can deal with this issue using the same type of regularization scheme discussed above. Figure 5.4C shows the solution obtained by regularizing the inversion. The "change in behaviour" for this problem occurs around $k = 4$. We see that the inverted $I(t)$ does indeed match the true $I(t)$ for most of S phase (Fig. 5.4C). To understand the decrease of the solution towards the end, we also plot the solution obtained using a higher threshold, $\tau = s_{j=20}^2$ (Fig. 5.4C). In contrast to the $\tau = s_4^2$ solution, the $\tau = s_{20}^2$ reconstruction increases throughout but is dominated by noise in the regime where the $\tau = s_4^2$ solution decreases. Thus, the decrease in the solution is a numerical artifact produced by the algorithm, in its attempt to eliminate large bin-to-bin fluctuations. Nonetheless, the algorithm provides a good reconstruction of $I(t)$ for most of S phase.

In summary, we have presented a method, following [127], to deconvolve an experimental FACS histogram into an ideal histogram. The histogram can be integrated to produce a scaled $f(t)$ via Eq. 5.1. Being the time-derivative of $f(t)$ (see Eq. 5.2), the FACS histogram also provides information on the domain density, since the rate of replication is determined by the number of replicated domains growing. (More precisely, $df/dt = 2vn_f$, where v is the fork velocity and n_f the density of replicating domains.) Lastly, the ideal FACS histogram can be used to infer the general shape of the underlying initiation rate $I(t)$.

5.3 Analysis of FACS-microarray data: time-averaged fit

In the previous section, we studied how to extract the spatially averaged replication fraction $f(t)$ and initiation rate $I(t)$ from FACS histograms. In the FACS-microarray technique, the cells selected by FACS are hybridized onto microarray chips. The result is a replication profile that reflects the temporally averaged replication fraction $f(x)$. In this section, we explore how much information we can extract from $f(x)$ via fitting.

The FACS-microarray replication profile $f(x)$ has contributions from cells that entered

S phase at different times. Formally,

$$f(x) = \int_{-\infty}^{\infty} f(x,t)\rho(t)dt, \tag{5.12}$$

where $f(x,t)$ is the replication fraction defined in Eq. 2.5, and $\rho(t)$ is the fraction of cells in the culture that have been in S phase for time t. For the analysis in this section, we assume $\rho(t)$ to be uniform between $[t_s, t_e]$ and 0 everywhere else. In practice, the time boundaries t_s and t_e can be estimated from the hard thresholds in the FACS histogram (Fig. 5.5). Here, we assume that t_s and t_e are known.

We expect a uniform $\rho(t)$ to be a reasonable estimate of the experimental distribution because, under normal growth conditions, the cells sampled by FACS are uniformly distributed in the cell cycle time. On the other hand, the experimental $\rho(t)$ will have a smooth boundary for the following reasons: In the FACS-microarray shown in Fig. 5.5, one considers a cell to be replicating if its DNA content is in between two hard thresholds. The hard thresholds in the experimental FACS histogram translate into smooth thresholds in the ideal DNA content because of the point-spread function. Also, because the replication process is stochastic, cells at different times in S phase can have the same DNA content. Thus, the soft thresholds in the ideal DNA content are further smoothed out. Overall, we expect that the shape of $\rho(t)$ to be flat topped with rounded edges (Fig. 5.5 bottom right).

Given $\rho(t)$, fitting is straightforward. One can simply generate $f(x,t)$ using the sigmoid model (SM) or the multiple-initiation model (MIM) in Chapters 3 and 4 and use Eq. 5.12 to calculate the averaged $f(x)$ to fit to the FACS-microarray data. Since this procedure is essentially the same as the analysis presented for time-course microarray, we first study the reliability of the FACS-microarray fit. In particular, we want to test the reliability of the FACS-microarray fits relative to the time-course-microarray fits.

5.3.1 Time-averaged fit vs. time-course fit

We first simulate a time-course microarray dataset and a FACS-microarray replication profile. For the time-course microarray, we simulate an ideal version of the dataset in [57], where "ideal" means that the experimental imperfections mentioned in Sec. 3.4 are absent. We simulate Chromosome XI of the budding yeast genome with the fork velocity extracted

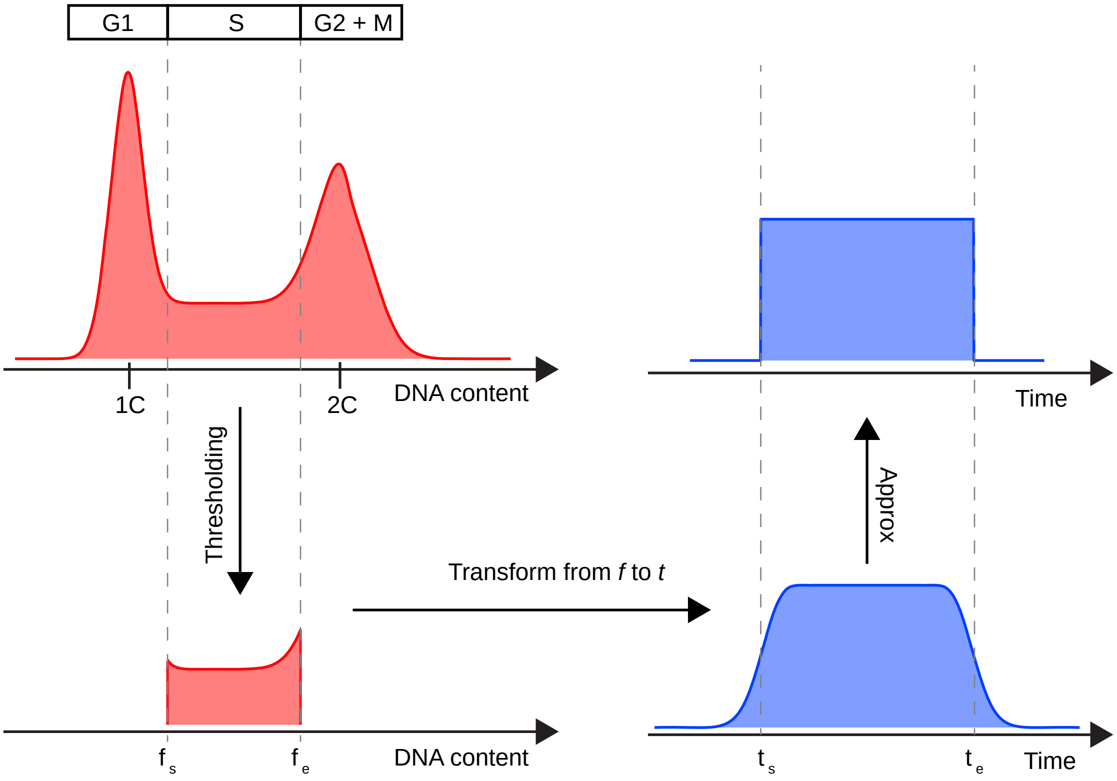


Figure 5.5: Two hard thresholds f_s and f_e are assigned to the FACS histogram for selecting cells that are in S phase. The truncated FACS histogram can be transformed into the corresponding distribution $\rho(t)$ that describes the fraction of cells in the culture that is at time t relative to the start of S phase. The distribution, bottom right, can be approximated by a uniform distribution with hard boundaries t_s and t_e, top right.

from SM (1.9 kb/min) and the final SM parameters tabulated in Supplementary Table I (Sec. 4.A.2). The dataset is composed of eight time points ranging from 10 to 45 minutes at 5-minute intervals. The spatial resolution is 1 kb. Gaussian measurement noise is added. The noise standard deviation for each time point is the estimate given in Table 3.1.

For the FACS-microarray simulation, the input parameters are the same as above. Of course, instead of eight synchronized time points, the FACS microarray generates only one averaged replication profile. The simulation uses a $\rho(t)$ that is uniform between $t_s = 10$ and $t_e = 45$ minutes relative to the start of S phase and 0 for other times. This corresponds to the scenario where the FACS uniformly samples cells from that part of S phase. Gaussian noise (standard deviation = 3.5% replication fraction) is added to the replication profile $f(x)$.

We fit both simulated datasets with the SM. In the fit, everything is known except for the fork velocity and the three SM parameters for each origin: origin position, median firing time $t_{1/2}$, and width of firing time t_w. The results are shown in Fig. 5.6. In general, the parameter uncertainties resulting from the FACS-microarray fit are larger than those from the time-course-microarray data. In particular, Fig. 5.6C shows that the uncertainties in t_w in FACS-microarray fit are usually larger than the parameter itself. This reflects the property of averaging; i.e., firing-time distributions with very different widths can have similar averages. In other words, the FACS-microarray data are nearly degenerate with respect to t_w variations. Figure 5.6D also supports this point: while the FACS-microarray t_w parameters are quite different from the input values, they still result in a profile that is similar to the theoretical one.

The large degeneracy in t_w suggests that the SM parameterization is not suitable for FACS-microarray data. From Fig. 5.6A–C, one sees that the data allow reliable estimates of roughly two parameters per origin: one for position and another for the firing time distribution or initiation rate. A reasonable choice for the latter is a linearly increasing initiation rate $I_i(t) = 2I_i t$, where $2I_i$ is the slope of the rate that characterizes the i^{th} origin. The main advantage of this choice is that it allows Eq. 5.12 to be calculated semi-analytically. From Eq. 2.5, one sees that the replication fraction $f(x,t) \propto \exp[-\int I_i(t)dt]$. Assuming that $\rho(t)$ is a uniform distribution with domain $[t_s, t_e]$, for $I_i(t) = 2I_i t$, Eq. 5.12 becomes

$$f(x) = 1 - \frac{1}{t_e - t_s} \int_{t_s}^{t_e} \exp\left[-\sum_{\text{all i}} I_i t^2\right] dt, \tag{5.13}$$

which can be calculated semi-analytically using error functions.

Furthermore, a linear initiation rate is biologically plausible. Figure 3.5 shows that the chromosome-averaged initiation rates in budding yeast are nearly linear in the time range (10–45 min) probed by the experiment. The same increasing trend is also observed in many species [38] and is a feature that makes the replication process robust (see Sec. 3.3.7). Interestingly, assigning each origin a slope $2I_i$ corresponds to a multiple-initiator model where every initiator follows the global initiation rate $I_o(t) = 2I_o t$ (see discussion around Eq. 4.4). We test the speed of this "linear MIM" on a FACS-microarray dataset probing

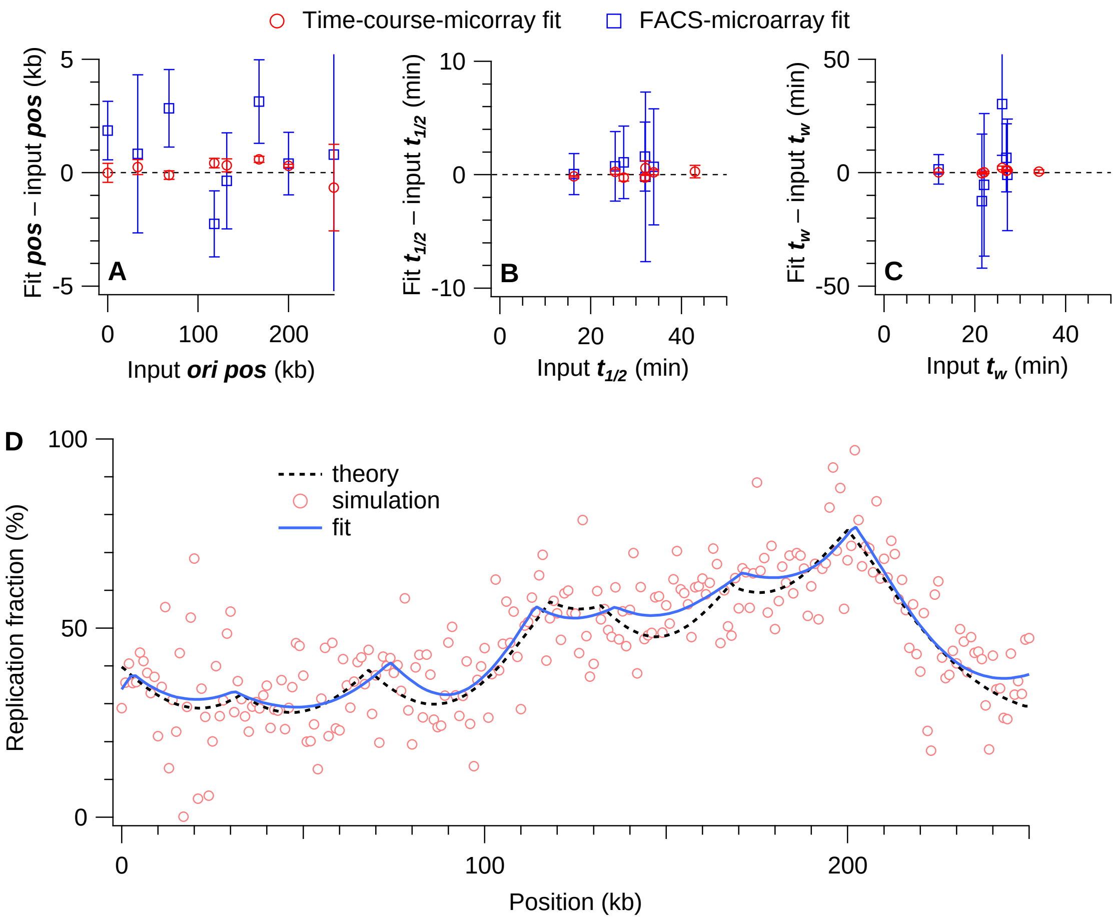


Figure 5.6: Comparison between the fits to time-course microarray and the fits to FACS-microarray data. **A–C**. The x-axis is the input parameter value used in the simulation. The y-axis is the difference between the fit value and the input value. Error bars are from uncertainties in the fit parameters. In B and C, the last FACS-microarray parameter is outside the displayed range and thus not shown. **D**. Fit to the simulated FACS-microarray data. The theoretical curve (dotted line) is calculated using the input values and overlaps the simulation curve before Gaussian noise (standard deviation = 3.5% replication fraction) is added. The fit to the noisy simulated data is plotted with a solid line.

budding yeast [53]. The fit took roughly 20 minutes, which is much shorter than the 10 hours needed for the time-course microarray.

5.4 Analysis of FACS-microarray data: non-parametric reconstruction

In our theory, the fundamental quantities characterizing the replication kinetics are the initiation rate $I(x,t)$ and the fork velocity v. From these two quantities, we derived the replication fraction $f(x,t)$ in Sec. 2.1.1. In the previous section, we built a forward model that connects these fundamental quantities to the FACS-microarray data that yield the spatially averaged replication fraction $f(t)$ and the temporally averaged replication fraction $f(x)$. Conceptually, the fitting methods used in Sec. 5.3 require solid prior information such as the number of origins. In this section, we explore another approach, similar to the one in Sec. 5.2.1, where the algorithm is non-parametric and is based on general considerations about the solution's structure. Our goal is to reconstruct $I(x,t)$ and $f(x,t)$ from $f(t)$ and $f(x)$ using such methods. In this analysis, we assume that the fork velocity v is known independently, in order to simplify the treatment.

Since the data are discrete, we again work with discretized quantities: $f(t) \rightarrow \mathbf{f^t}$, a column vector with n_t elements; $f(x) \rightarrow \mathbf{f^x}$, a column vector with n_x elements; $f(x,t) \rightarrow \mathbf{F^{xt}}$, a matrix with n_t-by-n_x elements; and $I(x,t) \rightarrow \mathbf{I^{xt}}$, also a matrix with n_t-by-n_x elements. We start with the matrix equation that connects $\mathbf{F^{xt}}$ to $\mathbf{f^t}$ and $\mathbf{f^x}$. For cleaner notation, we introduce an operator R that redimensions an n_t-by-n_x matrix into a column vector with $n_t n_x$ elements via

$$R(\mathbf{F^{xt}}) = \left[\mathrm{row}(\mathbf{F^{xt}},1), \mathrm{row}(\mathbf{F^{xt}},2), \cdots, \mathrm{row}(\mathbf{F^{xt}},\mathrm{n_t})\right]^{\mathrm{T}}, \tag{5.14}$$

where $\mathrm{row}(\mathbf{F^{xt}},\mathrm{i})$ extracts the i^{th} row from $\mathbf{F^{xt}}$ and $^{\mathrm{T}}$ denotes the transpose operator. Using this notation, the data are connected to $\mathbf{F^{xt}}$ via a simple matrix multiplication

$$\mathbf{d} = \begin{bmatrix} \mathbf{f^t} \\ \mathbf{f^x} \end{bmatrix} = \mathbf{A}R(\mathbf{F^{xt}}), \tag{5.15}$$

where $\mathbf{A}$ is an $(n_t + n_x)$-by-$(n_t n_x)$ averaging matrix that can be constructed in a straightforward manner.

As in Sec. 5.2, one could solve for $R(\mathbf{F^{xt}})$ in Eq. 5.15 in the least-squares sense by minimizing

$$\begin{aligned} J &= (\mathbf{A}R(\hat{\mathbf{F}}^{\mathbf{xt}}) - \mathbf{d})^{\mathrm{T}}(\mathbf{A}R(\hat{\mathbf{F}}^{\mathbf{xt}}) - \mathbf{d}) \\ &= \left\| \mathbf{A}R(\hat{\mathbf{F}}^{\mathbf{xt}}) - \mathbf{d} \right\|_2, \end{aligned} \tag{5.16}$$

where $\|\cdot\|_2$ is the ℓ_2, or Euclidean, norm, and $\hat{\mathbf{F}}^{\mathbf{xt}}$ is an estimate of $\mathbf{F}^{\mathbf{xt}}$. In contrast to the deconvolution problem in Sec. 5.2.1, the problem here is underdetermined, and the least-squares solver was not able to converge to even a noisy solution. To overcome this problem, we regularize J by adding the curvature term introduced in Sec. 5.2.1. Such a regularized cost function has two desirable features: First, minimizing the curvature leads to a specific structure and is thus much less degenerate than Eq. 5.16. Second, we expect biologically relevant $\mathbf{F}^{\mathbf{xt}}$ to be smooth, as has been observed in many organisms.

We construct a regularizing matrix $\mathbf{S}^{\mathbf{t}}$ via $\mathbf{S}^{\mathbf{t}}R(\mathbf{F}^{\mathbf{xt}}) = R(\mathbf{C}^{\mathbf{t}}\mathbf{F}^{\mathbf{xt}})$, where $\mathbf{C}^{\mathbf{t}}$ is the numerical second derivative matrix in Eq. 5.8. Since $\mathbf{F}^{\mathbf{xt}}$ is n_t-by-n_x, $\mathbf{C}^{\mathbf{t}}$ is n_t-by-n_t and measures the temporal curvature of $\mathbf{F}^{\mathbf{xt}}$. The matrix $\mathbf{S}^{\mathbf{t}}$ is then, by construction, $(n_t n_x)$-by-$(n_t n_x)$. In like manner, we also construct a regularizer $\mathbf{S}^{\mathbf{x}}$ for the spatial part via $\mathbf{S}^{\mathbf{x}}R(\mathbf{F}^{\mathbf{xt}}) = R(\mathbf{C}^{\mathbf{x}}\mathbf{F}^{\mathbf{xt}\mathrm{T}})$, where $\mathbf{C}^{\mathbf{x}}$ is n_x-by-n_x and measures the spatial curvature of $\mathbf{F}^{\mathbf{xt}}$. The matrix $\mathbf{S}^{\mathbf{x}}$ is also $(n_t n_x)$-by-$(n_t n_x)$.

Before adding the regularizers $\mathbf{S}^{\mathbf{t}}$ and $\mathbf{S}^{\mathbf{x}}$ to Eq. 5.16, we note that each element in the replication fraction matrix $\mathbf{F}^{\mathbf{xt}}$ must be between $[0, 1]$ to be biological. Put together, we have turned the reconstruction problem into a constrained optimization problem:

Minimize

$$J = \left\| \mathbf{A}R(\hat{\mathbf{F}}^{\mathbf{xt}}) - \mathbf{d} \right\|_2 + \lambda_1 \left\| \mathbf{S}^{\mathbf{t}}R(\hat{\mathbf{F}}^{\mathbf{xt}}) \right\|_2 + \lambda_2 \left\| \mathbf{S}^{\mathbf{x}}R(\hat{\mathbf{F}}^{\mathbf{xt}}) \right\|_2 \tag{5.17a}$$

subject to

$$0 \leq \hat{\mathbf{F}}^{\mathbf{xt}} \leq 1, \tag{5.17b}$$

where λ_1 and λ_2 are constants that adjust the weight of each factor in the objective function. We note that normal least-squares methods cannot handle the constraints in Eq. 5.17b. To solve this problem, we use `CVX`, a package for specifying and solving constrained convex programs [128, 129]*. The package runs in the MATLAB environment and includes solvers that use advanced interior-point methods [130]. Equation 5.17 is convex because both the objective function in Eq. 5.17a and the constraints in Eq. 5.17b are convex functions. Once $\hat{\mathbf{F}}^{\mathbf{xt}}$ is obtained, Eq. 2.9 can be used to extract $\hat{\mathbf{I}}^{\mathbf{xt}}$.

5.4.1 Reconstruction with ideal FACS-microarray data

To test the method outlined above, we simulate a 60-kb genome at 1-kb resolution from 0–20 min at 1-min resolution. The initiation rate is 0 everywhere except for the bins centred at $x = 10, 30, 50$ kb, as shown in Fig. 5.7A. The fork velocity is 1 kb/min. The simulated replication fraction $\mathbf{F}^{\mathbf{xt}}$ and the averages $\mathbf{f}^{\mathbf{t}}$ and $\mathbf{f}^{\mathbf{x}}$ are shown in Fig. 5.7C. The simulation procedures are described in Sec. 2.3. Our goal is to reconstruct $\mathbf{F}^{\mathbf{xt}}$ and $\mathbf{I}^{\mathbf{xt}}$, given the two "clues" $\mathbf{f}^{\mathbf{t}}$ and $\mathbf{f}^{\mathbf{x}}$. Figure 5.7B shows that the initiation rates have an increasing and a decreasing part. We use this form instead of the linear rate advertised in Sec. 5.3.1 because we want to test whether our reconstruction methods can capture both trends.

Solving Eq. 5.17 with $\lambda_1 = \lambda_2 = 1$, we first obtain $\hat{\mathbf{F}}^{\mathbf{xt}}$ (Fig. 5.8C). Given the fork velocity ($v = 1$ kb/min), we then use Eq. 2.9 to invert $\hat{\mathbf{I}}^{\mathbf{xt}}$ from $\hat{\mathbf{F}}^{\mathbf{xt}}$ (Fig. 5.8A and B). Our first observation is that while the reconstructed $\hat{\mathbf{f}}^{\mathbf{t}}$ and $\hat{\mathbf{f}}^{\mathbf{x}}$ agree with the simulation, the reconstructed initiation rates for the origins do not. Comparing Figs. 5.8B to 5.7B, we see that the reconstructed rates increase too slowly in the beginning and too quickly towards the end.

A related observation is that $\hat{\mathbf{I}}^{\mathbf{xt}}$ diverges at the latest time point (Fig. 5.8A). To explain this, we first recall that the inversion from $f(x,t)$ to $I(x,t)$ involves transforming $f(x,t)$ into $H(x,t) = -\ln[1 - f(x,t)]$ (see discussion around Eqs. 2.8 and 2.9). Here, we denote the discretized $H(x,t)$ by $\mathbf{H}^{\mathbf{xt}} = -\ln[1 - \mathbf{F}^{\mathbf{xt}}]$, where $\mathbf{H}^{\mathbf{xt}}$ is transformed element by element from $\mathbf{F}^{\mathbf{xt}}$. At late times, as $\hat{\mathbf{F}}^{\mathbf{xt}} \to 1$, $\hat{\mathbf{H}}^{\mathbf{xt}} \to \infty$. An investigation of the reconstructed $\hat{\mathbf{F}}^{\mathbf{xt}}$ and $\hat{\mathbf{H}}^{\mathbf{xt}}$ reveals that the divergence is exponential in time. Thus, $\hat{\mathbf{I}}^{\mathbf{xt}}$, being proportional to the second time derivative of $\hat{\mathbf{H}}^{\mathbf{xt}}$ (see Eqs. 2.8 and 2.9), also diverges. We speculate that the exponential divergence is related to the interior-point method which approximates the hard constraint of $\hat{\mathbf{F}}^{\mathbf{xt}} \leq 1$ in Eq. 5.17b by a steep but smooth log-barrier function [130]. In the future, it would be interesting to test this hypothesis using algorithms

*CVX is freely downloadable from `http://cvxr.com/cvx/`.

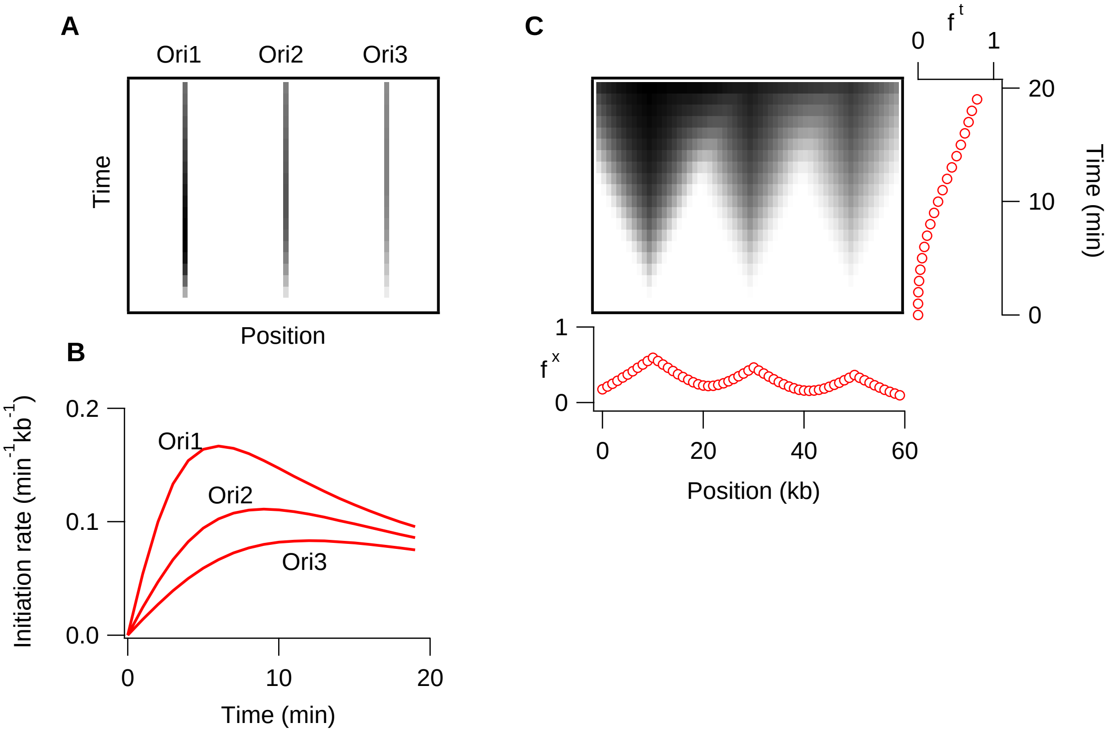


Figure 5.7: A simulation of FACS-microarray data. **A**. The $\mathbf{I^{xt}}$ matrix used to generate the simulation. The genome is 60 kb at 1-kb resolution. The simulation time spans 20 min at 1-min resolution. White is zero, and value increases with darkness. The three origins at 10 kb, 30 kb, and 50 kb, are labelled Ori1, Ori2, and Ori3, respectively. **B**. The initiation rates $I(t)$ for the three origins in A. **C**. The $\mathbf{F^{xt}}$ matrix. Same dimension as $\mathbf{I^{xt}}$. White is 0; black is 1. The spatial average $\mathbf{f^{x}}$ is shown at the bottom, and the temporal average $\mathbf{f^{t}}$ is shown on the right.

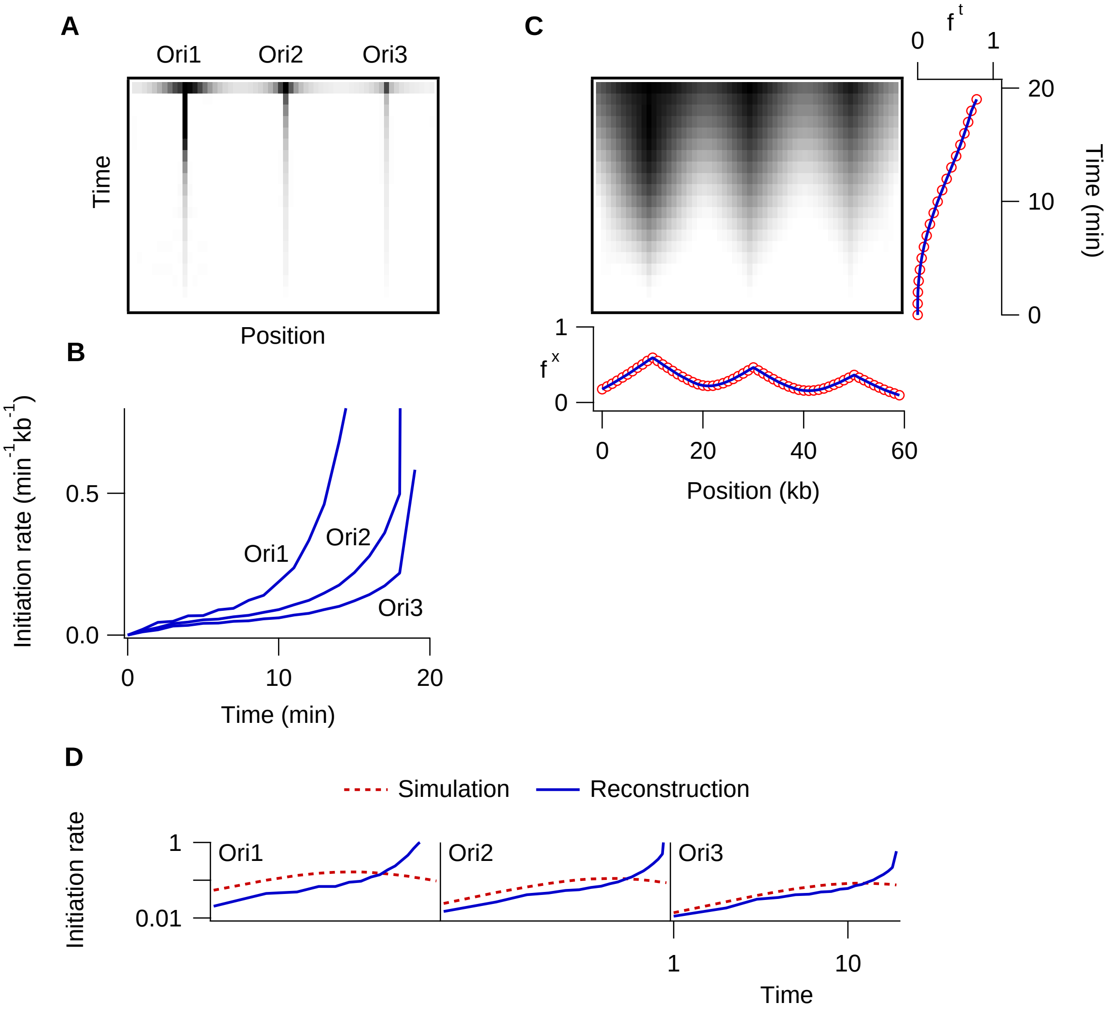


Figure 5.8: Reconstructing $\mathbf{F^{xt}}$ and $\mathbf{I^{xt}}$ from simulated FACS-microarray data in the ideal case. **A**. The reconstructed $\hat{\mathbf{I}}^{\mathbf{xt}}$ matrix. Same convention as in Fig. 5.7A. All values larger than 0.8 are shown as black for better display of the entire structure. **B**. The initiation rates $I(t)$ for the three origins in A. **C**. The reconstructed $\hat{\mathbf{F}}^{\mathbf{xt}}$ matrix. Same convention as in Fig. 5.7C. In the spatial and temporal average plots, markers are the input data (those from Fig. 5.7), and lines are the averages obtained from $\hat{\mathbf{F}}^{\mathbf{xt}}$. **D**. Comparison between the simulated and reconstructed origin initiation rates on a log-log scale.

that impose a hard constraint.

While the bias in our method towards an increasing rate is an issue, the main problem is that the data are consistent with a broad range of prior biases on the shape of the initiation rate. In other words, because many $\hat{\mathbf{I}}^{\mathbf{xt}}$ produce very similar $\mathbf{f^t}$ and $\mathbf{f^x}$, one has little sense of whether the reconstructed $\hat{\mathbf{I}}^{\mathbf{xt}}$ is similar to the true $\mathbf{I^{xt}}$. With information on only $\mathbf{f^t}$ and $\mathbf{f^x}$, the present method can reliably extract only one parameter (e.g., the relative efficiency) for each origin, in addition to the position. This is essentially the same conclusion that we obtained in Sec. 5.3.

On a more encouraging note, although the method does not capture the detailed shape of the initiation rates, it is surprisingly accurate in determining the origin positions. In particular, the reconstructed initiation rates concentrate at the correct origin positions to within 1 kb throughout time! Also, Fig. 5.8D shows that the linear increase of the reconstructed initiation rates at early times match that of the simulated rates.

5.4.2 Reconstruction with noisy FACS-microarray data

To test the method further, we add Gaussian noise to the input $\mathbf{f^t}$ and $\mathbf{f^x}$. We solve Eq. 5.17 with a range of values for λ_1 and λ_2; intuitively, large values of λ_1 and λ_2 suppress fluctuations in $\hat{\mathbf{F}}^{\mathbf{xt}}$. The solutions we obtain all exhibit two major problems. First, the reconstructed $\hat{\mathbf{I}}^{\mathbf{xt}}$ is not concentrated but has non-negligible values for most of the entries. This in turn leads to $\hat{\mathbf{F}}^{\mathbf{xt}}$, $\hat{\mathbf{f}}^{\mathbf{t}}$, and $\hat{\mathbf{f}}^{\mathbf{x}}$ that are smoother than supposed to be. Second, $\hat{\mathbf{I}}^{\mathbf{xt}}$ fluctuates between negative and positive values. Negative $\mathbf{I^{xt}}$ are not biological, as they correspond to unreplicating replicated domains. From these observations, we realize that the core problem with Eq. 5.17 is that the initiation rate $\hat{\mathbf{I}}^{\mathbf{xt}}$ is not constrained. In fact, because the replication fraction is a consequence of the initiation rate, it is best to work with $\mathbf{I^{xt}}$ in the reconstruction process rather than $\mathbf{F^{xt}}$.

Formulating reconstruction with $\mathbf{I^{xt}}$

To reformulate the problem in terms of $\mathbf{I^{xt}}$, we note that $\mathbf{H^{xt}}$, which equals $-\ln[1-\mathbf{F^{xt}}]$, is a linear combination of $\mathbf{I^{xt}}$. Thus, we define the matrix $\mathbf{P}$ via $R(\mathbf{H^{xt}}) = \mathbf{P}R(\mathbf{I^{xt}})$. Pictorially, just as a domain grows to fill in a “light cone” on the space-time graph that emanates from the initiation in Fig. 2.2, $\mathbf{P}$ propagates the initiation rate at (x, t) to fill the light cone that emanates from (x, t) with that particular rate (Fig. 5.9). Summing up all the light cones from each element in $\mathbf{I}^{\mathbf{xt}}$, one obtains $\mathbf{H}^{\mathbf{xt}}$ (Fig. 5.9). As an illustration, the scenario in Fig. 5.9 can be written explicitly as

$$\left[\begin{array}{cccc|cccc|cccc}
1 & 0 & \cdots & & & & & & & & \cdots & 0 \\
0 & 1 & 0 & \cdots & & & & & & & & \vdots \\
0 & 0 & 1 & 0 & \cdots & & & & & & & \\
0 & 0 & 0 & 1 & 0 & \cdots & & & & & & \\
\hline
1 & 1 & 0 & 0 & 1 & 0 & \cdots & & & & & \\
1 & 1 & 1 & 0 & 0 & 1 & 0 & \cdots & & & & \\
0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & \cdots & & & \\
0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & \cdots & & \\
\hline
1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & \cdots & \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & \cdots \\
1 & 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\
0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1
\end{array}\right]
\left[\begin{array}{c}
0 \\ I_1 \\ 0 \\ I_2 \\ \hline 0 \\ \vdots \\ \\ \\ \hline \\ \\ \vdots \\ 0
\end{array}\right]
=
\left[\begin{array}{c}
0 \\ I_1 \\ 0 \\ I_2 \\ \hline I_1 \\ I_1 \\ I_1 + I_2 \\ I_2 \\ \hline I_1 \\ I_1 + I_2 \\ I_1 + I_2 \\ I_1 + I_2
\end{array}\right]
\tag{5.18}$$

or in condensed form, $\mathbf{P}R(\mathbf{I}^{\mathbf{xt}}) = R(\mathbf{H}^{\mathbf{xt}})$. The elements indicated by the dots are all 0. We note that $\mathbf{P}$ is lower triangular because of causality. To explain this, we consider the first row of $\mathbf{P}$. If the 5^{th}–12^{th} elements of the first row of $\mathbf{P}$ were not zero, the 5^{th}–12^{th} elements of $R(\mathbf{I}^{\mathbf{xt}})$, which correspond to the initiation rates at later times, would propagate back in time to contribute to the first element of $R(\mathbf{H}^{\mathbf{xt}})$. Likewise, if the 2^{nd}–4^{th} elements of the first row of $\mathbf{P}$ were not zero, the 2^{nd}–4^{th} elements of $R(\mathbf{I}^{\mathbf{xt}})$, which correspond to the rate at positions 2–4, would influence position 1 in $R(\mathbf{H}^{\mathbf{xt}})$ in zero time. Applying the logic to every row, one sees that causality leads to a lower triangular structure. Although triangular matrices can be inverted efficiently, this feature is of only marginal benefit because the gradient-based optimization algorithm used does not involve inverting $\mathbf{P}$.

Having established the connection between $\hat{\mathbf{I}}^{\mathbf{xt}}$ and $\hat{\mathbf{F}}^{\mathbf{xt}}$ via $\mathbf{H}^{\mathbf{xt}}$ and $\mathbf{P}$, we can now

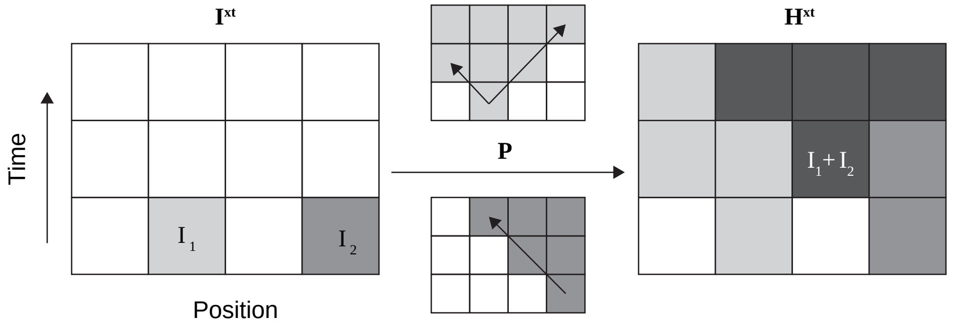


Figure 5.9: An illustration of the light-cone propagator. The matrix $\mathbf{I^{xt}}$ is non-zero at two space-time points with rate I_1 and I_2. The matrix $\mathbf{P}$ propagates the rates I_1 and I_2 forward in time and bidirectionally outward. The sum of the propagated rates is $\mathbf{H^{xt}}$.

consider the problem in terms of $\hat{\mathbf{I}}^{\mathbf{xt}}$. We propose solving the problem:

Minimize

$$J = \left\| \mathbf{A}\left(1 - e^{-\mathbf{P}R(\hat{\mathbf{I}}^{\mathbf{xt}})}\right) - \mathbf{d} \right\|_2 + \lambda_1 \left\| \mathbf{S}^{\mathbf{t}} R(\hat{\mathbf{I}}^{\mathbf{xt}}) \right\|_1 + \lambda_2 \left\| \mathbf{S}^{\mathbf{x}} R(\hat{\mathbf{I}}^{\mathbf{xt}}) \right\|_1 \tag{5.19a}$$

subject to

$$\hat{\mathbf{I}}^{\mathbf{xt}} \geq 0, \tag{5.19b}$$

where $\|\cdot\|_1$ is the ℓ_1 norm. Below, we discuss Eq. 5.19 term by term, including the choice of norms. The first term in the objective function J is simply a rewriting of the first term in Eq. 5.17a. However, the expression is now sigmoidal with respect to $\hat{\mathbf{I}}^{\mathbf{xt}}$ and non-convex. This means that we cannot use CVX to solve Eq. 5.19. One way to proceed is to employ algorithms that can handle nonlinear objective functions (e.g., [131, 132]). These algorithms usually involve approximating the constrained objective landscape locally as a convex problem and moving the local solution iteratively towards the local minimum. Instead of exploiting such nonlinear optimizers, we investigate the properties of Eq. 5.19, in particular, the effect of the ℓ_1 norm, in another way. As we will show later, the way involves formulating a problem that is similar to Eq. 5.19 but uses only linear functions of $\hat{\mathbf{F}}^{\mathbf{xt}}$ so that the objective function is convex. The application of suitable nonlinear programming algorithms to solve Eq. 5.19 is an interesting future topic.

Digression on ℓ_1 and ℓ_2 norms

Before explaining the second and third terms of Eq. 5.19, we first discuss the implication of the ℓ_1 norm. Compared to the ℓ_2 norm, the ℓ_1 norm induces sparse solutions; i.e., the quantity in $\|\cdot\|_1$ is exactly zero for many of its entries [133, 134]. To understand why, we follow [133] and consider the toy example,

$$\text{Minimize } \; J = \|\mathbf{Bx} - \mathbf{d}\|_2 + \lambda \|\mathbf{x}\|_1 \,, \tag{5.20}$$

where $\mathbf{B}$ is a 2-by-2 matrix, $\mathbf{x}$ is a column vector with entry x_1 and x_2, $\mathbf{d}$ is a 2-element column vector, and λ is a constant. Here, $\mathbf{B}$, $\mathbf{d}$, and λ are given, and x_1 and x_2 are unknown. Equivalently, the above unconstrained problem can be rewritten as a constrained problem:

$$\text{Minimize } \; J = \|\mathbf{Bx} - \mathbf{d}\|_2 \tag{5.21a}$$

$$\text{subject to } \; |x_1| + |x_2| < \tau, \tag{5.21b}$$

where τ can be calculated from λ and vice versa. Figure 5.10A shows the geometry of Eq. 5.21. The solution x_1 and x_2 occurs at the point where the contour levels of J in Eq. 5.21a first contacts the boundary of the constraint set by Eq. 5.21b. Figure 5.10B shows the same picture but for the constraint $x_1^2 + x_2^2 < \tau$. Note that replacing Eq. 5.21b with this constraint is equivalent to changing the second term in Eq. 5.20 from the ℓ_1 norm $\|\mathbf{x}\|_1$ to an ℓ_2 norm $\|\mathbf{x}\|_2$. Comparing Figs. 5.10A to 5.10B, one sees that the contact in the ℓ_1-norm picture sometimes occurs at the corners, where one of the $\mathbf{x}$ elements is exactly zero. In contrast, because the constraint in the ℓ_2-norm picture is round, the contact almost always occurs at points where both x_1 and x_2 are non-zero.

In higher dimensions, we expect the contact in the ℓ_1-norm picture to occur with increasing probability at "edges" of the constraint where some x values are exactly zero. For a hypercube of dimension n, the number of "faces" increases linearly with n but the number of "vertices" and "edges" exponentially with n. If the contact is at a face, all elements in $\mathbf{x}$ are non-zero; if not, some elements are exactly zero. As n increases, the number of vertices and edges quickly overwhelms the number of faces; thus, the probability that all elements are non-zero is minute. In contrast, because a hypersphere has no vertex or edge, the probability of having even one non-zero element is minute. This simple argument suggests that

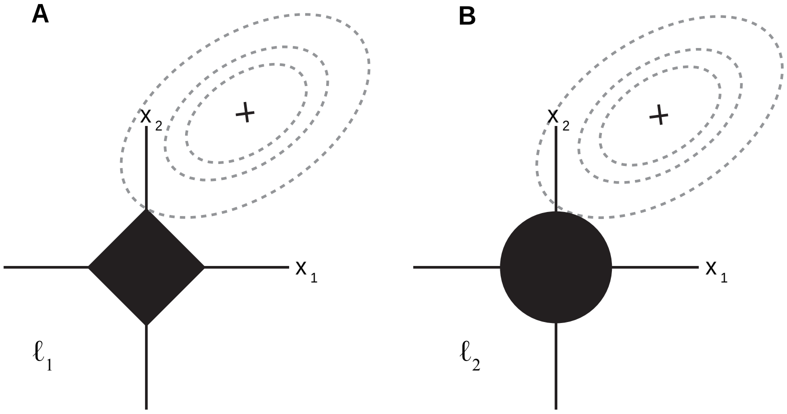


Figure 5.10: The geometry of ℓ_1 and ℓ_2 optimization. **A**. The solid diamond at the centre represents the constraint in Eq. 5.21b. The elliptical contours are the contour levels of J in Eq. 5.21a. The cross labels the solution of Eq. 5.21a without the constraint. The solution to the full problem in Eq. 5.21 is at the point where the contour level first contacts the solid diamond. **B**. Same as A, except that the constraint is now $x_1^2 + x_2^2 < \tau$. The illustration is adapted from Fig. 2 in [133].

the ℓ_1 norm is likely to induce sparse quantities, while the ℓ_2 norm is not.

Modified reconstruction formulation with $\mathbf{F}^{\mathbf{xt}}$

Having seen that ℓ_1 norm induces sparsity, we now explain the second and third terms of J in Eq. 5.19a. The rationale behind using the matrix $\mathbf{S}^{\mathbf{x}}$ and $\mathbf{S}^{\mathbf{t}}$ is the same—to penalize the spatiotemporal bin-to-bin fluctuations in $\mathbf{I}^{\mathbf{xt}}$—regardless of which of the norms is used. However, the use of ℓ_1 norm on $\mathbf{S}^{\mathbf{t}}R(\hat{\mathbf{I}}^{\mathbf{xt}})$ and $\mathbf{S}^{\mathbf{t}}R(\hat{\mathbf{I}}^{\mathbf{xt}})$ has the advantage that the reconstructed $\hat{\mathbf{I}}^{\mathbf{xt}}$ will have non-zero curvature for only few entries. Most often, zero curvature in $\hat{\mathbf{I}}^{\mathbf{xt}}$ also corresponds to zero initiation rate; thus, the use of the ℓ_1 norm here forces $\hat{\mathbf{I}}^{\mathbf{xt}}$ to be concentrated while regularizing rapid fluctuations. Lastly, Eq. 5.19b constrains $\hat{\mathbf{I}}^{\mathbf{xt}}$ to be ≥ 0. One can usually put an upper bound on $\hat{\mathbf{I}}^{\mathbf{xt}}$ as well, knowing that biologically plausible initiation rates cannot be too large.

The major difference between Eqs. 5.17 and 5.19 is the use of the ℓ_1 norm. To investigate the function of the ℓ_1 norm without resorting to nonlinear optimization software, we consider the convex problem*:

Minimize

$$J = \left\|\mathbf{A}R(\hat{\mathbf{F}}^{\mathbf{xt}}) - \mathbf{d}\right\|_2 + \lambda_1 \left\|\mathbf{S}^{\mathbf{t}}R(\hat{\mathbf{F}}^{\mathbf{xt}})\right\|_2 + \lambda_2 \left\|\mathbf{S}^{\mathbf{x}}R(\hat{\mathbf{F}}^{\mathbf{xt}})\right\|_2 + \lambda_3 \left\|\mathbf{S}^{\mathbf{t}}\mathbf{P}^{-1}R(\hat{\mathbf{F}}^{\mathbf{xt}})\right\|_1 \tag{5.22}$$

subject to

$$0 \leq \hat{\mathbf{F}}^{\mathbf{xt}} \leq 1$$

$$\mathbf{P}^{-1}R(\hat{\mathbf{F}}^{\mathbf{xt}}) \geq \alpha,$$

where λ_3 is a constant weight, and α is a constant lower bound. Compared to the original problem in Eq. 5.17, we have added in Eq. 5.22 the term $\left\|\mathbf{S}^{\mathbf{t}}\mathbf{P}^{-1}R(\hat{\mathbf{F}}^{\mathbf{xt}})\right\|_1$ to the objective function and $\mathbf{P}^{-1}R(\hat{\mathbf{F}}^{\mathbf{xt}}) \geq \alpha$ to the set of constraints. These two terms in Eq. 5.22 correspond to $\left\|\mathbf{S}^{\mathbf{t}}R(\hat{\mathbf{I}}^{\mathbf{xt}})\right\|_1$ and $\hat{\mathbf{I}}^{\mathbf{xt}} \geq 0$ in Eq. 5.19, respectively. (We leave out the corresponding $\|\mathbf{S}^{\mathbf{x}}\cdots\|_1$ term in this analysis so as to see more clearly the function of a single ℓ_1 norm.)

The rationale for formulating Eq. 5.22 is as follows: Although the use of $\hat{\mathbf{I}}^{\mathbf{xt}}$ in Eq. 5.19 is ideal for the reasons mentioned in the beginning of this section, it makes the problem non-convex. To keep the problem convex so that CVX can be used, we limit the objective function J to be a functional of only linear functions of $\hat{\mathbf{F}}^{\mathbf{xt}}$. At this point, we note that $\mathbf{P}^{-1}\hat{\mathbf{F}}^{\mathbf{xt}}$ and $\hat{\mathbf{I}}^{\mathbf{xt}} = \mathbf{P}^{-1}\hat{\mathbf{H}}^{\mathbf{xt}}$ have similar structures, implying that the solution obtained from minimizing $\left\|\mathbf{S}^{\mathbf{t}}\mathbf{P}^{-1}R(\hat{\mathbf{F}}^{\mathbf{xt}})\right\|_1$ should also be structurally similar to that obtained from minimizing $\left\|\mathbf{S}^{\mathbf{t}}R(\hat{\mathbf{I}}^{\mathbf{xt}})\right\|_1$. Thus, we propose solving Eq. 5.22 as a first step towards understanding the more ideal solution of Eq. 5.19.

In particular, since $\hat{\mathbf{H}}^{\mathbf{xt}}$ and $\hat{\mathbf{F}}^{\mathbf{xt}}$ are related to each other element by element, we expect the sparsity in $\mathbf{P}^{-1}\hat{\mathbf{H}}^{\mathbf{xt}}$ to be preserved in $\mathbf{P}^{-1}\hat{\mathbf{F}}^{\mathbf{xt}}$ for early times, before the origins start to passively replicate each other. To illustrate this, we consider a 5-by-5 example of $\mathbf{I}^{\mathbf{xt}} \rightarrow$

*It is known that $\|\mathbf{x}\|_p$ is convex for $p \geq 1$, where $\|\cdot\|_p$ is the p-norm. By extension, the p-norm of any linear function of $\mathbf{x}$ for $p \geq 1$ is also convex. Thus, both the ℓ_1 and ℓ_2 terms in Eq. 5.22 are convex, and CVX can be used to find the solution.

$\mathbf{H^{xt}}$:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 \end{bmatrix} \rightarrow \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \\ 6 & 3 & 1 & 0 & 0 \\ 10 & 6 & 3 & 1 & 0 \\ 15 & 10 & 6 & 3 & 1 \end{bmatrix}, \tag{5.23}$$

where → represents the transformation via the propagator $\mathbf{P}$. Time increases going down the rows of the matrices. The reverse transformation can be done via $\mathbf{P^{-1}}$ (the inverse of $\mathbf{P}$) which backpropagates the light cone in $\mathbf{H^{xt}}$ to a point source in $\mathbf{I^{xt}}$ according to Eq. 2.9. For the above case, which has fork velocity $v = 1$, the backpropagation in Eq. 2.9 simplifies to

$$I(x_r, t_s) = \frac{1}{2}\left[H(x_r, t_{s+1}) - H(x_{r-1}, t_s) + H(x_r, t_{s-1}) - H(x_{r+1}, t_s)\right], \tag{5.24}$$

where r and s are indices. Except for the boundary at the last row, we see that Eq. 5.24 recovers the sparsity in $\mathbf{I^{xt}}$ from $\mathbf{H^{xt}}$. Explicitly, this happens because $H(x_r, t_{s+1}) = H(x_{r-1}, t_s)$ and $H(x_r, t_{s-1}) = H(x_{r+1}, t_s)$ for places where $I(x_r, t_s) = 0$. Notice that because $\mathbf{H^{xt}}$ and $\mathbf{F^{xt}}$ are related element by element [via $\mathbf{F^{xt}} = 1 - \exp(\mathbf{H^{xt}})$], applying Eq. 5.24 to the corresponding $\mathbf{F^{xt}}$ also results in a sparse matrix, for the explicit reason mentioned above [i.e., $F(x_r, t_{s+1}) = F(x_{r-1}, t_s)$ and $F(x_r, t_{s-1}) = F(x_{r+1}, t_s)$]. This property no longer holds for $\mathbf{F^{xt}}$ at space-time points where multiple origins contribute to the replication fraction. When the sum contains more than one element, $\mathbf{F^{xt}} = 1 - \exp(-\sum \mathbf{H^{xt}}_i) \neq \sum[1 - \exp(-\mathbf{H^{xt}}_i)]$, where $\mathbf{H^{xt}}_i$ is the $\mathbf{H^{xt}}$ of the i^{th} origin. Thus, we expect that the $\mathbf{P^{-1}\hat{F}^{xt}}$ to be sparse until the time when origins start to interact.

Solving Eq. 5.22 using CVX with $\lambda_1 = \lambda_2 = 1$, $\lambda_3 = 5$, and $\alpha = -2$, we obtain Fig. 5.11. We chose these values through trial and error. The solutions are not very sensitive to the exact values of the λ and α: for instance, doubling all the values results in, on average, a 5% change in $\mathbf{\hat{f}^x}$ and $\mathbf{\hat{f}^t}$ and a 20% change in $\mathbf{\hat{I}^{xt}}$ around the origin positions. Figure 5.11C shows that the reconstructed $\mathbf{\hat{F}^{xt}}$, $\mathbf{\hat{f}^x}$, and $\mathbf{\hat{f}^t}$ are similar to the true inputs in Figure 5.7C. As argued above, this happens because the reconstructed $\mathbf{\hat{I}^{xt}}$ captures the general structure of the true $\mathbf{I^{xt}}$. Indeed, Fig. 5.11A shows that the reconstructed $\mathbf{\hat{I}^{xt}}$ is concentrated around the input origin positions, and Fig. 5.11B shows the same increasing trend as Fig. 5.8B. The similarity between Figs. 5.11 and 5.8 suggests that using the ℓ_1 norm, which constrains the solution to adopt the structure determined by the propagator, makes the reconstruction robust against noisy inputs.

A careful investigation of Fig. 5.11A and B reveals two issues. First, noise, though greatly reduced, still biases the estimate of the origin position and broadens its width, as shown in Fig. 5.11A and D. The noise also causes the three origins to be essentially indistinguishable (compare Fig. 5.11B to 5.8B). As one can see from Fig. 5.11C, the indistinguishable issue is expected because the noise level is comparable to the difference in peak heights (compare the noisy $\mathbf{f^x}$ in Fig. 5.11C to the $\mathbf{f^x}$ in Fig. 5.7C). Second, Fig. 5.11A shows traces of non-zero rates between the origins. As mentioned above, these traces are expected in the solution of Eq. 5.22 because the backpropagator $\mathbf{P^{-1}}$ acts on $\hat{\mathbf{F}}^{\mathbf{xt}}$ instead of on $\hat{\mathbf{H}}^{\mathbf{xt}}$. We expect this issue to disappear if one were to solve Eq. 5.19 for $\hat{\mathbf{I}}^{\mathbf{xt}}$ directly instead of solving for $\hat{\mathbf{F}}^{\mathbf{xt}}$ and inverting to $\hat{\mathbf{I}}^{\mathbf{xt}}$.

Another issue that can be more easily dealt with by working with $\hat{\mathbf{I}}^{\mathbf{xt}}$ rather than $\hat{\mathbf{F}}^{\mathbf{xt}}$ is the concentration of high initiation rates in $\hat{\mathbf{I}}^{\mathbf{xt}}$ at later times (Figs. 5.8A and 5.11A). As mentioned previously, the divergence happens because $\hat{\mathbf{F}}^{\mathbf{xt}} \to 1$ and $\hat{\mathbf{H}}^{\mathbf{xt}} \to \infty$ for late times. This is not an issue when $\hat{\mathbf{I}}^{\mathbf{xt}}$ is the fundamental variable, as the forward transformation, $\mathbf{F^{xt}} = 1 - \exp[-\mathbf{H^{xt}}]$, is well behaved. Furthermore, if one knows a priori that the initiation rate should decrease towards late S phase, as is the case of the simulation, one can directly constrain or bias $\hat{\mathbf{I}}^{\mathbf{xt}}$ towards the desired shape with a proper regularizing term.

In summary, we have developed a method to reconstruct the replication fraction $f(x,t)$ and $I(x,t)$ from $f(x)$ and $f(t)$—noise-corrupted spatial and temporal projections of $f(x,t)$ given by the FACS-microarray data. The method is based on solving a constrained convex optimization problem that constrains the solution to follow the structure set by the propagator $\mathbf{P}$ (or, equivalently, the backpropagator in Eq. 2.9). Using this method, we were able to reconstruct an estimate of $f(x,t)$ and $I(x,t)$ that capture many of the biological relevant features. In particular, we were able to reconstruct regions of concentrated initiation rates in $I(x,t)$ that correspond to origins.

Lastly, we note that the ideas presented in this section also apply to time-course microarray data. We mentioned in Sec. 2.1.1 two issues with the direct application of the inversion formula Eq. 2.9 to time-course microarray data: First, numerical differentiation amplifies noise in the data. Second, the temporal and spatial resolution of the data are often

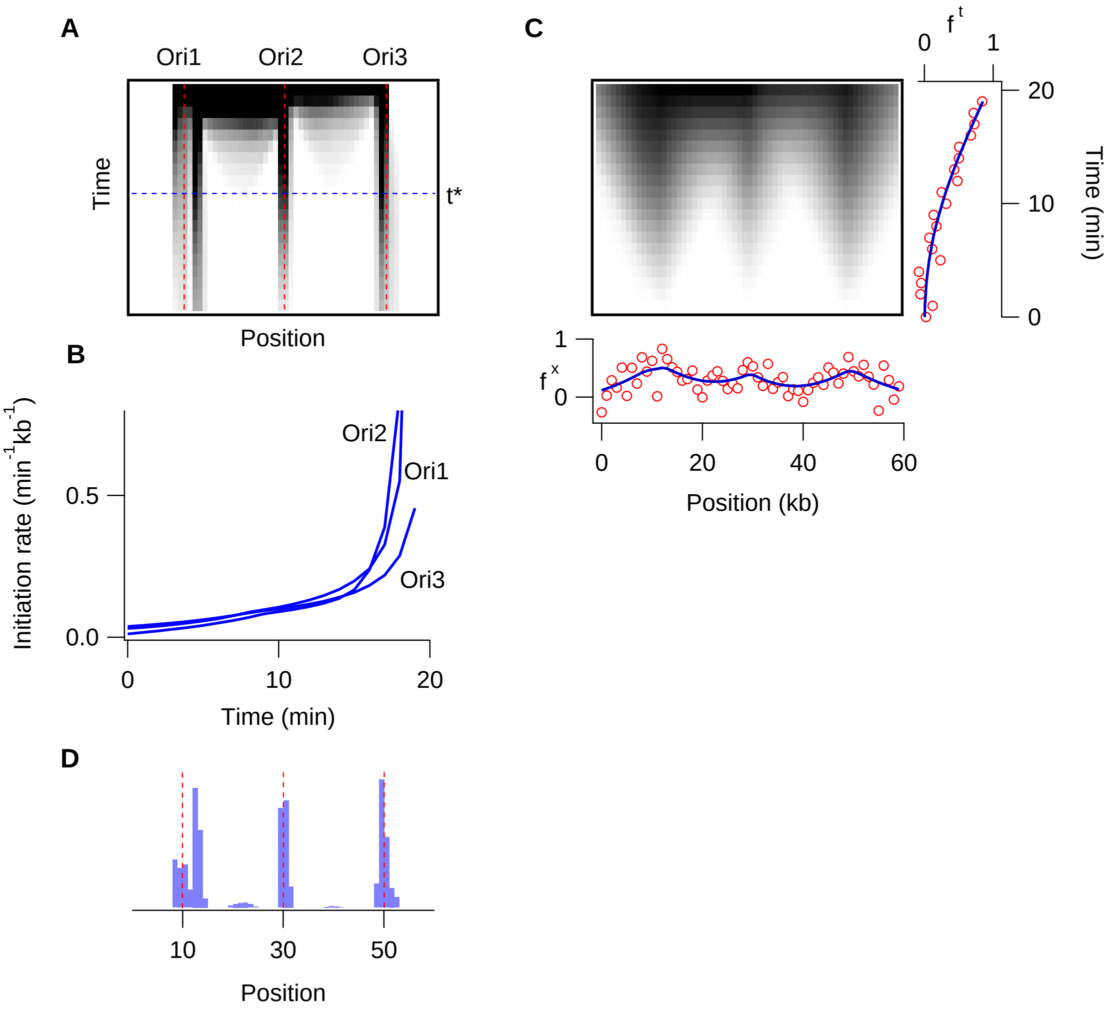


Figure 5.11: Reconstructing $\mathbf{F}^{\mathbf{xt}}$ and $\mathbf{I}^{\mathbf{xt}}$ from simulated FACS-microarray data with noise. **A**. The reconstructed $\hat{\mathbf{I}}^{\mathbf{xt}}$ matrix. Same convention as in Fig. 5.7A. White is 0; black is 0.05. The limited range is chosen to make the features between the origins more apparent. The vertical dotted lines mark the true positions of Ori1, Ori2, and Ori3. The slice of $\mathbf{I}^{\mathbf{xt}}$ marked by the horizontal line labelled t^* is shown in D. **B**. The initiation rates $I(t)$ for the three origins in A. Each curve is calculated as the sum over ± 2.5 kb of the marked origin positions in A. Similar to Fig. 5.8D, the rates increase linearly at earlier times. **C**. The reconstructed $\hat{\mathbf{F}}^{\mathbf{xt}}$ matrix. Same convention as in Fig. 5.7C. In the spatial and temporal average plots, markers are the input data, and lines are the averages from the reconstructed $\hat{\mathbf{F}}^{\mathbf{xt}}$. The input data are generated by adding Gaussian noise to $\mathbf{f}^{\mathbf{x}}$ and $\mathbf{f}^{\mathbf{t}}$ in Fig. 5.7C. **D**. A slice of $\mathbf{I}^{\mathbf{xt}}$ at time t^*. Vertical dotted lines are the true origin positions, as in A.

on different scales. Both problems can potentially be overcome by the method presented here. Replacing $\mathbf{d}$ with the time-course microarray data and $\mathbf{A}$ with the appropriate matrix, one can obtain a smooth $\hat{\mathbf{F}}^{\mathbf{xt}}$ (or $\hat{\mathbf{I}}^{\mathbf{xt}}$) with the appropriate regularizers (such as those in Eq. 5.22 or 5.19), even when $\mathbf{d}$ is noisy. To deal with the second issue, one can simply choose the dimensions of $\hat{\mathbf{F}}^{\mathbf{xt}}$ (or $\hat{\mathbf{I}}^{\mathbf{xt}}$) so that the spatial and temporal resolution are as desired. This often amounts to using a higher temporal resolution than the data offer. In this sense, the method is also a smoothing and interpolating algorithm that accounts for the structure of the replication fraction set by the propagator $\mathbf{P}$. We also note that in the case of time-course microarray, the method can be applied sequentially to short stretches of the genome and is can thus be easily scaled up to analyze replication in human.

Chapter 6

The Random-Completion Problem in Frog Embryos

In this chapter, we shift focus to replication in frog embryos, where DNA synthesis initiates stochastically in time and space. (Note that this is different from the case of budding yeast discussed in Chapter 3 and 4, where licensing is sequence specific.) Stochastic initiation implies fluctuations in the time to complete replication. These variations may lead to cell death if replication takes longer than the cell cycle time ($\approx$ 25 min for embryos). Surprisingly, although the typical replication time is about 20 min, *in vivo* experiments show that replication fails to complete at most once in 300 times. How is replication timing accurately controlled despite the stochasticity?

Biologists have proposed two solutions to this "random-completion problem." The first solution uses regularly spaced origins, while the second uses randomly located origins but increases their rate of initiation as S phase proceeds. We used the theory developed in Sec. 2.2 to investigate this problem. We argue that the biologists' second solution to the problem is not only consistent with experiment but also nearly optimizes the use of replicative proteins. We also show that spatial regularity in origin placement does not alter significantly the distribution of replication times and, thus, is not needed for the control of replication timing. In Sec. 1.4, we mentioned that this thesis has three themes: 1) developing models for eukaryotic replication, 2) applying the models to experiments, and 3) understanding the replication timing control in eukaryotes. This chapter, in illustrating how modelling can clarify and quantify the control of replication-completion time in eukaryotes in the presence of spatiotemporal stochasticity, is an example of themes 1 and 3.

6.1 Introduction

Many of the DNA-combing experiments mentioned in Sec. 1.3 have been done on embryos of the South African clawed frog, *Xenopus laevis* [61, 135, 54]. In contrast to the replication program in budding yeast investigated in Chapters 3 and 4, studies of the kinetics of replication in *Xenopus* embryos revealed a particularly interesting scenario where the replication program is stochastic not only in time but also in space [61, 18]. As we will discuss below, this spatiotemporal stochasticity has important implications for the development of embryonic cells.

In previous work, we mapped the stochastic replication process onto a one-dimensional nucleation-and-growth process and modelled the detailed kinetics of replication seen in DNA-combing experiments [54, 33, 34]. In [136], Bechhoefer and Marshall extended the model to quantitatively address a generalized version of the "random-completion problem," which asks how cells can accurately control the replication completion time despite the stochasticity. In this chapter, we present and extend the work in [136] further to investigate the idea that cells regulate the replication process in order to minimize their use of cell "resources" and to explore the effects of spatial regularity on the placement of origins.

6.1.1 The random-completion problem

Replication in *Xenopus* embryos is interesting because the process is stochastic yet the replication-completion times are tightly controlled. After fertilization, a *Xenopus* embryo undergoes 12 rounds of synchronous, uninterrupted, and abbreviated cell cycles (lacking G1 and G2 phases), whose durations are strictly controlled by biochemical processes that are independent of replication [1, 99]. In contrast to the case of most somatic cells, these embryonic cells lack an efficient S/M checkpoint to delay entrance into mitosis for unusually slow replication [137]. Nonetheless, in each embryonic cell cycle, roughly 3 billion basepairs of DNA are replicated in a 20-min S phase followed by a 5-min mitosis (M) phase at 23°C [138][†]. If replication is not completed before the end of mitosis, the cell suffers a "mitotic catastrophe" where the chromosomes break, eventually leading to cell death [1, 140, 141]. (See Sec. 6.3.1 for more discussion.) In replicating the lengthy genome, $\mathcal{O}(10^6)$ potential origins are licensed, without sequence specificity, and initiated stochastically throughout S phase [18, 54, 142, 108, 143]. One might expect that this spatiotemporal stochasticity leads to large fluctuations in replication times, which would result in frequent mitotic catastrophes. However, experiments imply that such catastrophic events for *Xenopus* embryos happen less than once in 300 instances (see Sec. 6.3.1). This means that despite the stochasticity in licensing and initiations, *Xenopus* embryos can tightly control the duration of S phase, in order to meet the 25-min "deadline" imposed by the cell-cycle duration.

Laskey was the first to ask whether non-sequence-specific licensing might lead to incomplete replication [144]. Specifically, he assumed that origins in embryonic cells initiate at the start of S phase. He then noted that if the origins were licensed at random, they would have an exponential distribution of separations. With the estimates of the average inter-origin spacing and fork velocity known at that time, one would expect a few very large gaps. The extra time needed to replicate the gaps would then imply a replication time longer than the known duration of S phase. Even though some details have changed, biologists still have such a paradox in mind when they refer to the random-completion problem [138].

In older references of replication (e.g., [145]), it was assumed implicitly that the potential origins are associated with ORCs. The estimated number of ORCs per nucleus in *Xenopus* embryos is about 3.5×10^5 (1 ORC per 8 kb) [142]. Positioning these ORCs randomly on the genome (non-sequence specificity assumption), one would find many gaps that cannot be replicated in time [18, 138]. However, more recent experiments revealed that initiations coincide with the MCM2-7 rings and that each ORC loads 20–40 copies of MCM2-7 [143, 108]. Using a pair of MCM rings as a potential origin, one then expects about 3.5–7 $\times 10^6$ potential origins per nucleus (1.9±0.6 potential origins/kb). Assuming that the potential origins are free to move and are uniformly distributed along the DNA, this

[†] The durations of the embryonic cell cycle depend on temperature. For this chapter, we take the cell cycle time to be 25 min at 23°C [139]. A typical duration of S phase used is $\approx$ 20 min [138, 99]. Longer times (25 min for S phase and 30 min for the cell cycle) have been observed at 20°C [139].

density implies that there is negligible chance of having a gap that is too large to replicate in time. Although a large excess of potential origins seems to resolve the problem, the actual distribution of these origins are not known. There is evidence that potential origins can cluster together, effectively reducing the average density [138]. In addition, experiments also show that potential origins initiate throughout S phase in a stochastic manner [54]. These effects will be discussed later in the chapter.

Over the years, biologists have proposed two qualitative scenarios to address this random-completion paradox. The first scenario, the "regular-spacing model," incorporates mechanisms that regularize the placement of potential origins despite the non-sequence specificity to suppress large inter-origin gaps [99]. The second scenario, the "origin-redundancy model," uses a large excess of randomly licensed potential origins and initiates them with increasing probability throughout S phase [54, 99, 146]. Experimentally, the observed replication kinetics in *Xenopus* are compatible with the origin-redundancy model, but there is also evidence for some regularity in the origin spacings [62, 138, 147].

In Chapter 2, we formulated the random-completion problem in a more general way. In particular, we investigated not only the possibility of replication completion but also the probability of completion (fluctuations in completion time). Using the theory developed in Chapter 2, we investigate here how cells control the replication time despite the non-sequence-specific placement and stochastic initiation of potential origins. As we shall see, the fluctuations in the replication times can be reduced arbitrarily if one allows an unrestricted number of initiations. As an extreme example, having an infinite number of initiations at time t^* implies that replication would always finish at t^*. Thus, an even more general formulation of the random-completion problem is to ask how reliability in timing control can be achieved with a reasonable or "optimal" use of resources in the cell. Of course, the terms "reasonable", "optimal", and "resources" must be carefully defined.

This chapter is organized as follows: In Sec. 6.2, we describe how we model the replication kinetics in *Xenopus* embryos and show how replication-completion time can be controlled despite the stochasticity. In Sec. 6.3, we use the model to extract parameters relevant to replication completion from *in vivo* and *in vitro* experiments. In Sec. 6.4, we compare the extracted *in vivo* "replication strategy" with the strategy that optimizes the activity of replication forks. In Sec. 6.5, we explore the effect of spatial ordering on the replication time via a variant of the regular-spacing model. We summarize our findings in Sec. 6.6.

6.2 Modelling replication completion

We start with a brief review of the model, which is described in detail in Sec. 2.2. Our model of the replication kinetics has three elements: initiation, growth, and coalescence of replicated domains (Fig. 2.2). In Sec. 2.2, we used a time-dependent initiation rate $I(t)$ and a constant fork velocity v to describe the replication kinetics in *Xenopus* embryos. Our model suggests that there is a one-to-one mapping between initiations and coalescences and that the last coalescence marks replication completion (Fig. 2.2). We were thus able to express implicitly the distribution of replication-completion times (or the end-time distribution) as a function of $I(t)$, v, and the genome length L. In particular, we showed that the end-time distribution is an extreme-value distribution, the Gumbel distribution, and depends on only two parameters, the mode t^* and the width β of the distribution defined in Sec. 2.2.

In a previous analysis, Herrick *et al.* extracted a bi-linear $I(t)$ from an *in vitro* DNA-combing experiment on *Xenopus* embryos [54]. In order to address the random-completion problem which is *in vivo*, we will transform the bi-linear initiation rate *in vitro* (I_{vitro}) to obtain a scaled bi-linear initiation rate *in vivo* (I_{vivo}) in Sec. 6.3. For ease of calculation, in parts of the paper, we approximate both bi-linear functions with power-law functions. Both initiation rates turn out to be approximately quadratic ($I_{vitro} \sim t^{2.62}$ and $I_{vivo} \sim t^{2.45}$).

The use of a bi-linear or quadratic form implies that the initiation rate increases throughout S phase. At the time that [54] was written, there was little evidence that suggested otherwise. Although the data in [54] show that $I(t) \to 0$ toward the end of S phase, the errors on the decreasing part are large, and the decrease was neglected [33, 136]. However, a more recent repeat of the original experiment has shown that the decrease is not an artifact of poor statistics but represents a true feature of the replication process [139, 38, 37]. To address this issue, we ran simulations and found that with the decrease seen in [38, 37], the mode of the end-time distribution was delayed by $\approx$ 0.3% and the width increased by 15%. These quantities approximately translate into an S phase that is prolonged by $\approx$ 0.5 min. Since this difference is small compared to the overall duration of S phase (20 min), we will use the simpler increasing $I(t)$ throughout the chapter.

Along with the bi-linear initiation rate, Herrick *et al.* extracted a constant fork velocity $v \approx 0.6$ kb/min [54]. In a more recent experiment, it was shown that the fork velocity *in vitro* decreases from 1.1 to 0.3 kb/min [148]. To test whether replacing the observed decreasing fork velocity with its average is a valid approximation, we simulated the two cases. We found that the tails of the coalescence distributions for the two cases agree to $\approx 1\%$, indicating that the two cases also have similar end-time distributions.

It is not surprising that the initiation rate and fork velocity are not as simple as we modelled them to be. In addition to the temporal variations that we mentioned above, there are also spatial variations and correlations [18, 19, 62, 138, 147, 149]. In Sec. 6.5, we will investigate the implications of spatial regularity in origin spacing by simulation and show that realistic regularity does not alter the end-time distribution. Overall, these simulations suggest that an increasing time-dependent $I(t)$ and a constant v are reasonable approximations.

6.2.1 Control of replication-completion time

As a first step toward resolving the random-completion problem, we consider the end-time distributions produced by different initiation rates. We use a power-law initiation rate [$I(t) = I_n t^n$, with I_n a constant and n the power] because it captures I_{vitro}, covers a wide range of scenarios, and can be treated analytically. We also examine an alternative δ-function form [$I(t) = I_\delta \delta(t)$, with I_δ a constant], where all potential origins initiate at the start of S phase, as one might expect this to be the scenario that minimizes replication time and its fluctuations. (In the early literature on DNA replication, biologists assumed this scenario to be true [144].) The most relevant results are the relationship among the replication kinetic parameters (I_δ, I_n, n, v, and L) and the end-time parameters (t^* and β) in Eqs. 2.18 and 2.23–2.24.

From the theory developed in Sec. 2.2, we infer two heuristic principles for controlling the end-time distribution: the first narrows the width, whereas the second adjusts the mode. To explore how the width β depends on the initiation form [$\delta(t)$ and t^n], we simulate the replication process by choosing the prefactors I_δ and I_n, using Eq. 2.21 so that the typical replication-completion time and fork velocity match the values inferred from *in vitro* experiments: $t^* = 38$ min and $v = 0.6$ kb/min. (As we will discuss in Sec. 6.3, replication *in vitro* is slower than *in vivo*.) The $t^* = 38$ min is obtained by simulating the end-time distribution with the bi-linear $I_{vitro}(t)$. The *Xenopus* genome length L is

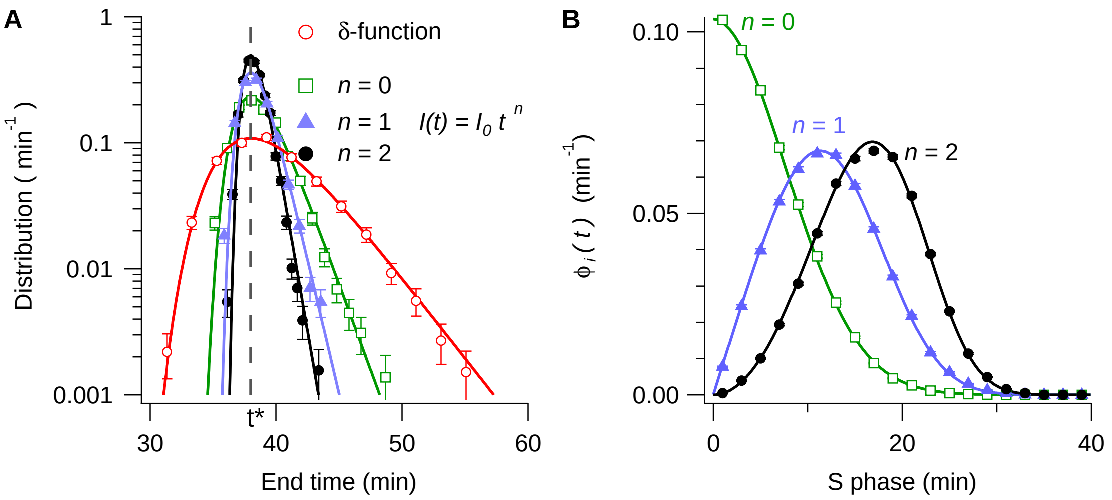


Figure 6.1: **A**. The end-time distribution with fixed mode $t^* = 38$ min. Markers are the results of the Monte Carlo simulations. Each distribution is estimated from 3,000 end times. The "δ-function" corresponds to initiating all potential origins simultaneously at $t = 0$ min. The $n = 0$, 1, 2 cases correspond to constant, linearly increasing, and quadratically increasing initiation rates, respectively. Solid lines are Gumbel distributions with t^* and β calculated according to Eqs. 2.21–2.22. There are no fit parameters. **B**. Initiation distribution $\phi_i(t)$ defined in Eq. 2.17 for $n = 0$, 1, 2. Parameter values correspond to those in **A**. Error bars are smaller than marker size. Solid lines are calculated from Eq. 2.17. Again, there are no fit parameters.

3.07×10^6 kb [150].

The result shown in Fig. 6.1A is perhaps counterintuitive: Initiating all origins in the beginning of S phase, which corresponds to a δ-function $I(t)$, gives rise to the broadest distribution. Initiating origins throughout S phase narrows the end-time distribution. The narrowing is more pronounced as the power-law exponent n increases. These observations can be explained by Eq. 2.22, which states that the width is inversely proportional to the average density of potential origins. The physical interpretation is that having fewer potential origin sites leads to more variation in the spacing between potential origins. This variation in turn induces fluctuations in the largest spacings between initiated origins, which widens the end-time distribution. In this light, Fig. 6.1A shows that when t^* is fixed, the δ-function case uses the fewest potential origins and thus produces the widest distribution. In contrast, a large power-law exponent n implies the use of many potential origins and thus produces a narrow distribution. In summary, the first heuristic principle is that the end-time distribution can be narrowed arbitrarily by increasing the number of potential origins in the system.

The second principle is that given an excess of potential origins, cells can initiate origins progressively throughout S phase instead of all at once to lower the consumption of resources while still controlling the typical replication time. In S phase, activators and polymerases are recyclable proteins; i.e., they can be reused once they are liberated from the DNA [151]. Progressive initiation then allows a copy of the replicative protein to be used multiple times. Compared to initiating all origins at once, this strategy requires fewer copies of replicative machinery and thus saves resources. This notion of minimizing the required replication resources is further discussed in Sec. 6.4.

Figure 6.1B shows that increasing the exponent n results in the “holding back” of more and more initiations until later in S phase. Comparing Figs. 6.1B with 6.1A, one finds that “holding back” initiations corresponds to narrowing the end-time distribution. Although many potential origins are passively replicated and thus never initiate, the timing of replication can still be accurately controlled, as initiations now occur in the “needed places.” Since the probability of initiation inside a hole (an unreplicated region) is proportional to the size of the hole, the held-back initiations are more likely to occur in large holes. This filling mechanism is made efficient by increasing $I(t)$ toward the end of S phase so that any remaining large holes are increasingly likely to be covered.

One subtle point of the origin-redundancy scenario is that although the potential origins are licensed at random, the spacings between initiated origins form a distribution $\rho_i(s)$ with a non-zero mode that contrasts with the exponential distribution of spacings between potential origins. An example of the $\rho_i(s)$ is shown later in Sec. 6.5. In earlier literature, before experiments showed that initiations can take place throughout S phase, biologists believed that all potential origins initiate at the start of S phase. In this δ-function case, the distribution of the inter-potential-origin spacing is the same as that of the spacing between fired origins (inter-origin spacing). As mentioned previously, a completely random placement leads to an exponential distribution. Thus, the mode of the inter-origin distribution $\rho_i(s)$ is at zero spacing ($s = 0$): there are many very small gaps balanced by occasional large ones. By contrast, in a scenario with an increasing $I(t)$, a peak will arise in $\rho_i(s)$ because closely spaced potential origins are not likely to all initiate but be passively replicated by a nearby initiation. This passive replication effect suppresses the likelihood of having small inter-origin spacings and thus creates a non-zero mode value in the spacing distribution. One should be careful not to confuse the two distributions.

In conclusion, we have shown that a large excess of potential origins suppresses fluctuations in the size of inter-potential-origin gaps, while the strategy of holding back initiations allows control of the typical replication time. These control mechanisms are also "open loop" in that they do not require any information about the replication state of the cell. In the next section, we review what is known experimentally about DNA replication kinetics in *Xenopus* embryos, in light of the analysis we have just presented.

6.3 Analysis of replication experiments in frog embryos

In the previous section and in Sec. 2.2, we showed that given an initiation rate and a fork velocity, one can find the associated end-time distribution using extreme-value theory. In this section, we review what is known experimentally about these quantities in *Xenopus* embryos. There have been two classes of experiments: *in vivo*, where limited work has been done [140, 1, 141], and *in vitro*, where rather more detailed studies have been performed on cell-free extracts [61, 135, 138, 54]. Typically, embryo replication *in vivo* takes about 20 minutes of the (abbreviated) 25-minute cell cycle [99]. As we discuss below, *in vivo* experiments imply that replication "failure"—incomplete replication by the end of the cell cycle—is very unlikely, occurring less than once in about 300 instances. The *in vitro* experiments on cell-free extracts give more detailed information about the replication process, including an estimate of the *in-vitro* initiation rate $I_{vitro}(t)$. However, the typical replication time *in vitro* is about 38 min, not 20 min, and it is not obvious how one can apply the results learned from the experiments *in vitro* to the living system. Below, we propose a way to transform $I_{vitro}(t)$ into an estimate of the *in vivo* initiation rate $I_{vivo}(t)$ that satisfies the failure probability of the *in vivo* system.

6.3.1 *In-vivo* experiments on cell death

A low replication-failure rate is remarkable because *Xenopus* embryos lack an efficient S/M checkpoint to delay cell cycle progression when replication is incomplete [99]. If chromosomes separate before replication is complete, cells suffer "mitotic catastrophe," which leads to apoptosis [140]. Thus, a low failure rate in embryonic cells implies that replication timing is precisely controlled by the initiation rate and fork velocity. Mathematically, we can test whether an initiation rate is realistic by calculating the rate of mitotic catastrophe F it implies. To evaluate F, we first choose a time t^{**} at which mitotic catastrophe occurs if replication is not fully completed. Then,

$$F \equiv \int_{t^{**}}^{\infty} \phi_e(t) dt \;=\; 1 - \Phi_e(t^{**}) \; . \tag{6.1}$$

As a first step in estimating F, we identify t^{**} with the cell cycle time ($\approx$ 25 min). Our identification is justified by observations that imply that replication can continue throughout mitosis, if needed [140]. Thus, even if the bulk of replication is completed before entering mitosis, small parts of the genome may continue to replicate, essentially until the cell totally divides. However, if unreplicated regions remain after the cell finishes dividing, the two daughter cells will inherit fragmented chromosomes.

Having identified t^{**}, we estimate F using data from an experiment on DNA damage in embryos [1, 141]. In [1], Hensey and Gautier found that cells with massive DNA damage (induced by radiation) will continue to divide through 10 generations. Then, at the onset of gastrulation, which occurs between the 10^{th} and 11^{th} cleavages, an embryo triggers a developmental checkpoint that activates programmed cell death. The role of cell death is to eliminate abnormal cells before entering the next phase of development, where the embryo's morphology is constructed via cell migration. In Hensey and Gautier's study, abnormal cells were detected using TUNEL staining, a technique for detecting DNA fragmentation in cells. In a later work investigating the spatial-temporal distribution of cell deaths in *Xenopus* embryos, they reported that, at gastrulation, 67% of 237 embryos, each containing 1024 cells, had more than 5 TUNEL-stained cells [141]. We can estimate F from the above observations using a simple model based on the following four elements:

1. All cells divide; each produces two cells.

2. If a cell has an abnormal chromosome, all its progeny are abnormal because replication can at best duplicate the parent's chromosome.

3. Along with abnormal cell division and other factors, failure to replicate all DNA

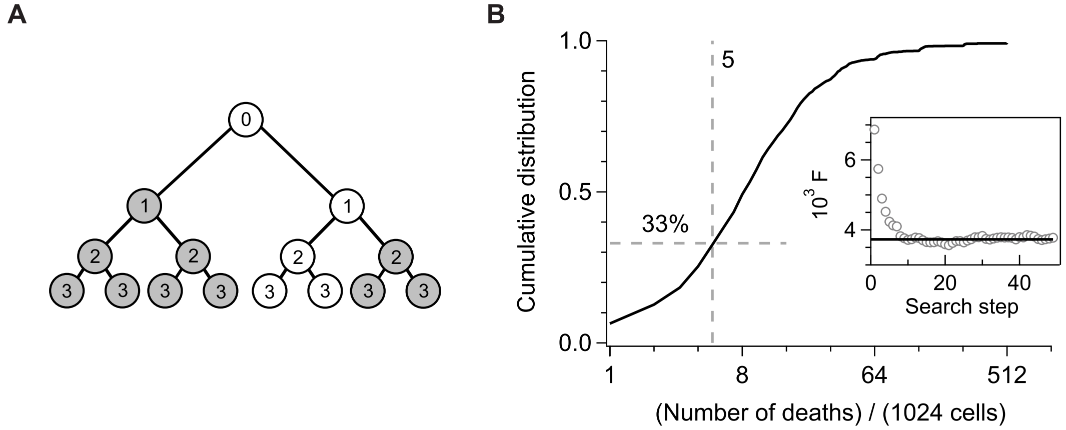


Figure 6.2: A branching model for cell proliferation. **A**. Schematic diagram of the simple model described in the text. Open circles represent normal proliferating cells, while filled circles represent abnormal cells. The numbers indicate the round of cleavage. Once a cell fails to replicate properly, all its progeny will be abnormal. **B**. Cumulative distribution of the number of dead cells at gastrulation (between cleavage 10 and 11) generated using Monte Carlo simulation. The distribution satisfies the constraint that 33% of the embryos have 5 or fewer abnormal cells. The inset shows the convergence of the gradient search to $F = 3.73 \pm 0.01 \times 10^{-3}$. The average and standard deviation of the mean are computed over the last 40 values of the gradient search.

before the end of a cell cycle results in chromosome breakages and leads to apoptosis at gastrulation.

4. All normal cells in all rounds of cleavage have the same probability F of becoming abnormal because of incomplete replication.

A schematic depiction of our model is shown in Fig. 6.2A.

The above model can be described by a standard Galton-Watson (GW) branching process [152], where the number of proliferating progeny generated by a normal cell is an independent and identically-distributed random variable. GW processes obey recursion relations that can be solved analytically using probability generating functions; however, the solution in our case is too complex to be helpful. We thus turned to numerical analysis.

We used Monte Carlo methods to simulate the branching process outlined above. Each embryo, after going through 10 rounds of division, contains m abnormal cells that commit apoptosis before the 11^{th} division. Simulating N embryos results in a distribution of number of deaths. We then compare the value of the cumulative distribution at 5 death events with the reported likelihood, which states that 33% of the time, there are 5 or fewer dead cells in 1024 cells [141]. Figure 6.2B shows the cumulative distribution that matches the reported numbers. To find F, we used a gradient-based method for finding roots of stochastic functions [153]. In this case, the input is the failure rate F, and the function evaluates the number and likelihood of deaths via a Monte Carlo simulation of the branching process of 237 embryos. We found that the numbers reported in [141] imply $F = 3.73 \pm 0.01 \times 10^{-3}$ (Fig. 6.2B inset)[†]. Since replication failure is only one of the factors that contribute to cell death, the F ($\approx$ 1 in 300) inferred is an upper bound to the replication failure rate.

Comparing Eq. 6.1 with the standard cumulative Gumbel distribution given by the integral of Eq. 2.18, one can relate the quantities t^{**} and F to the Gumbel parameters via

$$t^{**} = t^* - \beta(t^*) \ln \left[\ln \left(\frac{1}{1-F} \right) \right] \ . \tag{6.2}$$

For $F \ll 1$, the expression simplifies to $t^{**} \approx t^* - \beta(t^*) \ln(F)$, which implies that the end time is insensitive to the exact value of F: an order-of-magnitude estimate suffices.

6.3.2 Connecting replication *in vitro* to duplication failure *in vivo*

As discussed above, the most detailed experiments on replication in *Xenopus* have been conducted on cell-free egg extracts. In previous work [54], Herrick *et al.* modelled a DNA-combing experiment on such an *in vitro* system and inferred the time-dependent initiation rate $I_{vitro}(t)$ (approximately quadratic as discussed in Sec. 6.2), a fork velocity of 0.6 kb/min (averaged over S phase [148]), and a typical replication time t^* of 38 min. In contrast, the typical replication time in living embryos is only 20 min. While it is generally believed that DNA replication in the two settings occurs in a similar way, the overall duration of S phase is an obvious difference that must be reconciled. We thus have a dilemma: the known replication parameters, v and $I(t)$, are extracted from *in vitro* experiments while

[†] In [136], we estimated a failure rate $F \approx 10^{-4}$ using a simple model that neglected the complications due to the branching process. Since t^{**} depends on the logarithm of F from Eq. 6.2, the factor of 30 between the two estimates of F results only in a roughly 1 min shift in the t^{**} of the end-time distribution and does not alter the qualitative conclusion of the previous work.

the failure rate F is derived from observations of cells *in vivo*. Is it possible to "transpose" the results from the *in vitro* experiments to the *in vivo* setting? Although any such transformation is obviously speculative, we propose here a simple way that is consistent with known experimental results.

We hypothesize that, except for the fork velocity, replication is unaltered between the *in vitro* and *in vivo* systems. The subtlety is that there are several conceivable interpretations of "unaltered" replication. One could keep $I_{vitro}(t)$ the same; however, this is not reasonable in that the dramatic increase in $I_{vitro}(t)$, at $t \approx 17.4$ min of the bi-linear function, would be moved from the midpoint of replication to the end [54]. Alternatively, one could express the initiation rate in terms of the fraction of replication; i.e., $I = I(f)$, and preserve this function. In this case, one would need a fork velocity of about 2.2 kb/min to produce the extracted *in vivo* failure rate. Although this is a reasonable fork speed in systems such as the *Drosophila* embryo, it is about twice the maximum fork speed observed in *Xenopus* embryonic replication *in vitro* [148]. The third possibility is to preserve the maximum number of simultaneously active replication forks. Intuitively, this is plausible as each replication fork implies the existence of a large set of associated proteins. The maximum fork density then gives the minimum number of copies of each protein set required. Thus, we are in effect assuming that the number of replicative proteins remains the same in both cases.

The simplest way to preserve fork usage is to rescale the density of forks active at time t,

$$n_f(t) = \frac{1}{2v}\frac{df}{dt} = g(t)e^{-2vh(t)} , \tag{6.3}$$

linearly in time so that

$$n_f^{vivo}\left(\frac{t}{t_{scale}}\right) = n_f^{vitro}\left(\frac{t}{t^*_{vitro}}\right) , \tag{6.4}$$

where $t^*_{vitro} \approx 38$ min and t_{scale} is chosen so that $t^{**} = 25$ min and $F = 3.73 \pm 0.01 \times 10^{-3}$. We found that the *in vitro* fork usage is preserved by using the rescaling $I_{vivo}(t/t_{scale}) \sim 2I_{vitro}(t/t^*_{vitro})$ and $v = 1.030 \pm 0.001$ kb/min (Fig. 6.3)[†]. The error on v is a consequence

[†] In extracting the initiation rate I_{vitro}, Herrick *et al.* also extracted a "starting-time" distribution [54]. As discussed at the end of Sec. 2.2.3, the starting-time asynchrony partially captures the spatial variations in the true initiation rate and should be accounted for. We showed that such variations shift the mode of the end-time distribution. The starting-time distribution extracted in [54] is a normal distribution with standard deviation ≈ 6 min. Using Eq. 2.27, we estimate that t^*_{vitro} increases by roughly 5%. Since the change is

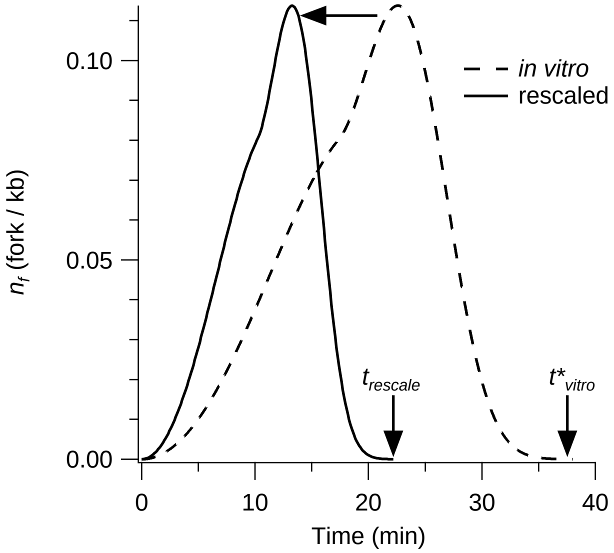


Figure 6.3: Density of simultaneously active replication forks throughout S phase, $n_f(t)$. The dotted curve corresponds to the *in vitro* fork usage while the solid curve is the rescaled fork usage that satisfies the constraints $t^{**} = 25$ min and $F = 0.00373$. The rescaled $n_f(t)$ is generated using $I_{vivo}(t/t^*_{vivo}) \sim 2I_{vitro}(t/t^*_{vitro})$ and $v = 1.030$ kb/min.

of the uncertainty in F.

Using the transformed $I_{vivo}(t)$, we estimate from $g_{vivo}(t^*)$ the lower bound of the potential origin density to be 1.2 potential origins/kb (PO/kb). This lower bound is consistent with the experimentally estimated average density of 1.9±0.6 PO/kb mentioned in Sec. 6.1.1. The velocity we infer also has a significant interpretation. As noted previously, Marheineke and Hyrien found that the fork velocity *in vitro* is not constant but decreases linearly from about 1.1 kb/min to 0.3 kb/min at the end of S phase [148]. The decrease in fork velocity suggests that *in vitro* replication progressively depletes rate-limiting factors (e.g., the nucleotides) throughout S phase. We suggest that our extracted $v \approx 1$ kb/min means that *in-vivo* systems are able to maintain the concentration of rate-limiting factors, perhaps by the factors' diffusing into the nuclear membrane [154] to maintain a roughly constant fork velocity throughout S phase. In summary, by preserving the rescaled version of the fork usage rate *in vitro*, we have transformed $I_{vitro}(t)$ into an $I_{vivo}(t)$ that satisfies the *in-vivo* failure rate and results in reasonable replication parameters.

small, the rescaling arguement presented here still holds.

6.4 Replication completion and the optimization of fork activity

The random-completion problem mentioned in Sec. 6.1 can be quantitatively recast into a problem of searching for an initiation rate that produces the *in-vivo* failure rate constraint in Eq. 6.1. In Fig. 6.4A, we show that any initiation form with the proper prefactor can satisfy the constraint on the integral of the end-time distribution, including the transformed *in vivo* initiation rate. Can we then understand why *Xenopus* embryos adopt the roughly quadratic $I(t)$ and not some other function of time?

To explore this question, we calculate for the different cases of $I(t)$ the maximum number of simultaneously active forks. Figure 6.4B shows that initiating all origins at the start of S phase [setting $I(t) \sim \delta(t)$] requires a higher maximum than a modestly increasing $I(t)$. At the other extreme, a rapidly increasing $I(t)$ (high exponent n) also requires many copies of replicative machinery because the bulk of replication is delayed and needs many forks close to the end of S phase to finish the replication on time. Thus, intuitively, one expects that an intermediate $I(t)$ that increases throughout S phase—but not too much—would minimize the use of replicative proteins. Figure 6.4B hints that the *in vivo* initiation rate derived from *in vitro* experiments may be close to such an optimal $I(t)$, as the number of resources required by $I_{vivo}(t)$ is close to the minimum of the power-law case.

The three resources modelled explicitly are potential origins, activators, and replication forks. It is not immediately clear which replication resources should be optimized. In general, the metabolic costs of expressing genes and making proteins are assumed to be non-rate-limiting factors. On the other hand, it is plausible that the cell minimizes the "complexity" of the replication process to minimize topological problems caused by simultaneously active replication forks, thus reducing the chance of unfaithful replication. Thus, in our optimization analysis, we ignore the metabolic costs of having a large number of potential origins and propose that the maximum number of simultaneously active forks is minimized. We argued previously that the maximum of $n_f(t)$ gives the minimum number of copies of the proteins required for DNA synthesis. Moreover, since the unwinding and synthesis of DNA at the forks create torsional stress on the chromosomes, minimizing the number of active forks would minimize the complexity of the chromosome topology, which may help maintain replication fidelity [155]. For these reasons, the maximum number of

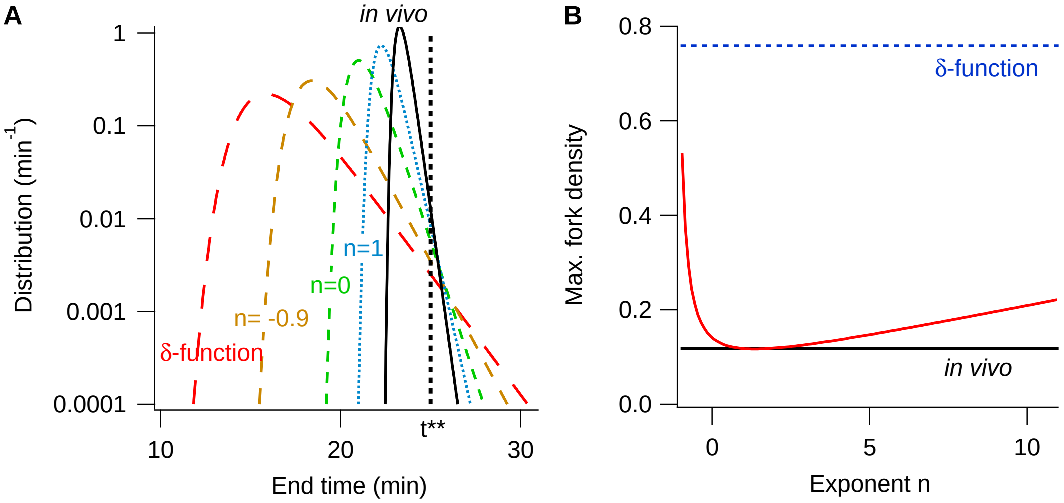


Figure 6.4: **A**. Replication end-time distribution with t^{**} fixed to be 25 min and $F = 0.00373$. Similar to Fig. 6.1A, the width decreases with an increase in the exponent n. **B**. Typical maximum number of simultaneously active forks. The curve is obtained from extracting the maximum value of $n_f(t)$ for different exponents n.

active forks is a plausible limiting factor for replication. Below, we calculate the optimal $I(t)$ and compare it with $I_{vivo}(t)$.

The number of forks active at time t is given by $n_f(t) = 2g(t)\exp[-2vh(t)]$. One can find the $I(t)$ that optimizes the maximum of $n_f(t)$ by minimizing

$$n_{max}[I(t)] = \lim_{p\to\infty}\left[\int_0^{t^{**}} n_f\,[I(t)]^p\ dt\right]^{1/p} . \tag{6.5}$$

This is a common analytic method to optimize the maximum of a function [156]. The trick is to analytically calculate the Euler-Lagrange equations for finite p and then take the limit $p \to \infty$, where the contribution of the maximum dominates the integrand. The associated Euler-Lagrange equation is

$$\ddot{h}(t) = 2v\dot{h}^2(t) , \tag{6.6}$$

where we recall that $\ddot{h}(t) = I(t)$ and $\dot{h}(t) = g(t)$. Note that Eq. 6.6 is independent of p, suggesting that the optimal $n_f(t)$ does not have a peak. Solving Eq. 6.6 subject to the boundary condition that the replication fraction be 0 at $t = 0$ [i.e., $h(0) = 0$] and 1 at

$t = t^{**}$, we obtain

$$I_{opt}(t) = \frac{1}{2vt^{**}} \left[\delta(t) + \frac{1}{t^{**}} \frac{1}{(1 - t/t^{**})^2} \right] \; . \tag{6.7}$$

Inserting the result from Eq. 6.7 into Eq. 6.3, one sees that $n_f(t) = 1/vt^{**}$ is constant throughout S phase and is about three times smaller than the maximum number of simultaneously active forks *in vivo* (Fig. 6.5C). Intuitively, the most efficient strategy is to use all the forks all the time. In terms of $I_{opt}(t)$, this optimal solution, like $I_{vivo}(t)$, increases slowly at first, then grows rapidly toward the end of S phase (Fig. 6.5B). However, the diverging initiation probability at $t \to t^{**}$ implies that this initiation rate is unphysical. In effect, a constant fork density implies that when the protein complexes associated with two coalescing forks are liberated, they instantly find and attach to unreplicated parts of the chromosome. It also implies that at the end of S phase, all the replication forks would be active on a vanishingly small length of unreplicated genome. Both implications are unrealistic.

To find a more realistic solution, we tamed the behaviour of the initiation rate for $t \to t^{**}$ by adding a constraint. A natural constraint to impose is that the failure rate *in vivo* be satisfied[†]. The diverging initiation rate at $t = t^{**}$ in Eq. 6.7 means that the replication always finishes exactly at t^{**} and that the failure rate is zero. Therefore, having a non-zero failure rate would force the initiation rate to be non-divergent. This constraint is also consistent with the idea that the replication process is shaped by the evolutionary pressure of survival. The new optimization quantity is then

$$J\left[I(t)\right] = \max\left\{n_f\left[I(t)\right]\right\} + \lambda\left\{F\left[I(t)\right] - F_{vivo}\right\} , \tag{6.8}$$

where the first term is the maximum of the fork density, and the second term is a penalty function that increases J for $F \neq F_{vivo}$. The strength of the penalty is set by the Lagrange multiplier λ. The time associated with F is $t^{**} = 25$ min throughout this section.

Substituting Eq. 6.5 into the first term of Eq. 6.8 and applying the method of variational calculus, we obtained an integro-differential equation that is difficult to solve analytically because the gradient of Eq. 6.5 is highly nonlinear and because F depends on t^*, which is

[†] Since a diverging initiation rate requires infinitely many activators and initiators to find each other in an infinitesimal amount of time, another natural constraint is that the number of replicative proteins and search time be finite, as modelled in [37]. This is an interesting topic for future analysis.

not readily expressible in terms of the basic replication parameters $I(t)$ and v. For these reasons, we turned to a gradient-approximation numerical method called finite difference stochastic approximation (FDSA) [153]. Although this search method is used for stochastic functions (as the name suggests), the method is just as suitable for deterministic functions. The basic concept is that the gradient of a function, which encodes the steepest-decent direction toward a local minimum, can be approximated by a finite difference of the function. The advantage of this method is that we can replace the complicated evaluation of the variation $\delta J[I(t)]$ by the easily calculable difference $J[I + \Delta I] - J[I - \Delta I]$.

Figure 6.5 shows the results of the FDSA search. We perform FDSA under several different conditions, with the initial search function being $I_{vivo}(t)$. First, we investigate the case where the optimization objective J is simply $\max\{n_f\}$, with no constraint or boundary condition [except $n_f(t) > 0$]. The markers in Fig. 6.5A shows that the optimal solution lingers near $\max\{n_f\} = 0.05$ (slow decrease in J) and then goes to the global minimum (zero). In the transient regime (search step between 50 to 100), the fork density evolves from a bell curve to a constant, which is the form of the calculated optimal solution. For search step > 100, the fork density (a constant) decreases to zero if no constraint is imposed. This zero solution corresponds to the case where no initiation or replication occurs. However, when the boundary condition used in the calculation (replication finished at t^{**}) is imposed, the FDSA algorithm indeed finds the $n_f(t) = 1/vt^{**}$ optimal solution (data not shown).

The second search was implemented following Eq. 6.8, where the constraint in F is added. Figure 6.5C shows that the fork solution is no longer a constant because the tail needs to decrease to satisfy $F = F_{vivo}$. The corresponding effect on the $I(t)$ is a decrease toward the end of S phase (Fig. 6.5B). The $I(t)$ behaves otherwise as predicted by Eq. 6.7 for most of S phase—a δ-function at the beginning followed by a rate that increases sharply at the end of S phase. Interestingly, the mechanism of spreading out the fork density to minimize the maximum fork usage seen in the analytical calculation is still present here, as shown by the plateau at early S phase (Fig. 6.5C).

In the third search, in addition to Eq. 6.8, we impose that there be no burst of initiation at the beginning of S phase [$g(0) = 0$], as seen in experiments. Figure 6.5C shows that with the addition of each constraint, the maximum of the fork density increases toward the *in vivo* value. Furthermore, besides satisfying the constraints and boundary conditions, the

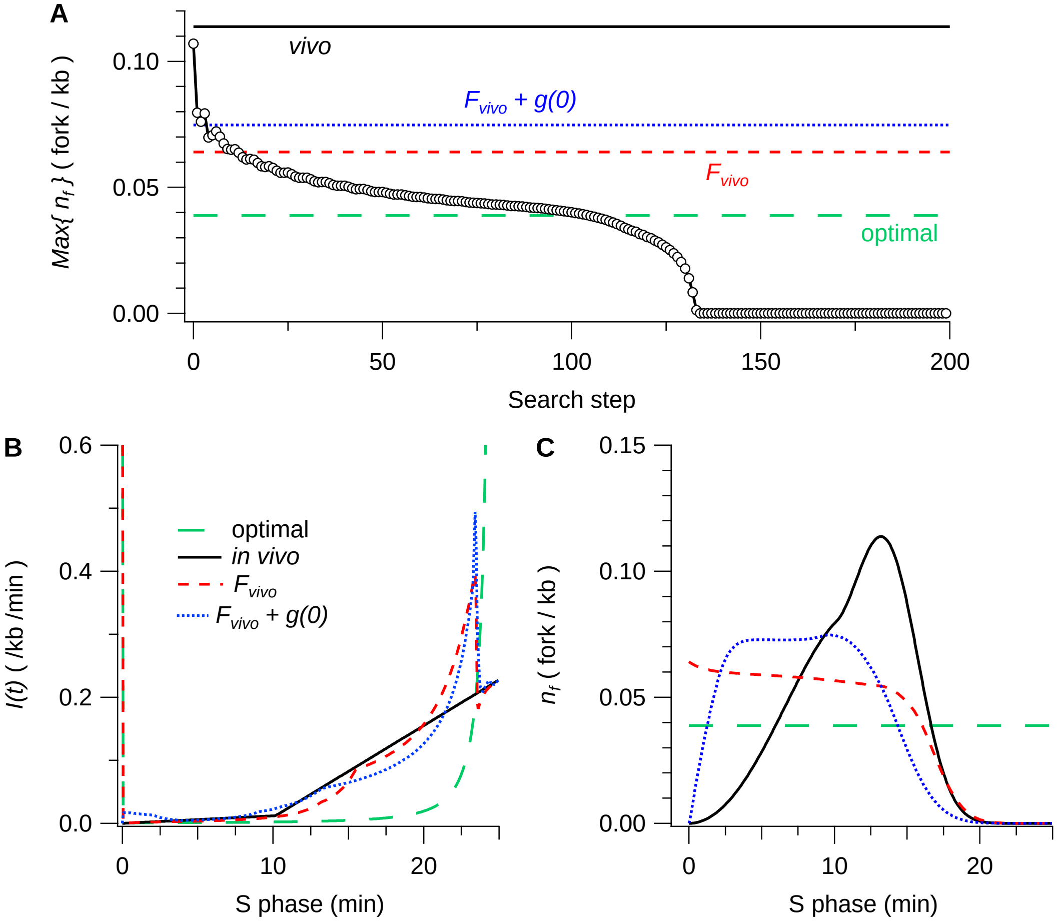


Figure 6.5: Results of a numerical search for optimal initiation rates under various constraints. The label "*vivo*" corresponds to the *in vivo* case; "optimal" corresponds to optimizing maximum fork density with no constraint (corresponds to Eq. 6.7); "F_{vivo}" corresponds to optimization with the constraint that the failure rate be equal the F_{vivo} extracted in Sec. 6.3.1; "$F_{vivo} + g(0)$" corresponds to optimization with the constraint of F_{vivo} and the constraint that $g(0) = 0$. **A**. Finite difference stochastic approximation search. The markers show the search for the case of minimizing the $\max\{n_f\}$ with no constraint and no boundary condition. The horizontal lines are the maximum fork density for different search conditions. **B**. Initiation rate $I(t)$. The I_{vivo} shown is in the bi-linear form, following [54]. **C**. Fork density $n_f(t)$. Line types correspond to those in **B**.

fork density profiles show a common feature of forming as lengthy a plateau as possible to minimize the maximum. The resulting $I(t)$ with this additional constraint is qualitatively similar to I_{vivo} (Fig. 6.5B).

There are still some differences between the result of the third search and n_f^{vivo}. In particular, the optimal fork solution increases much faster at the beginning of S phase than n_f^{vivo} does to spread out the fork activities. Minimizing the maximum number of initiations also leads to the same feature: a fast initial increase in the initiation activities followed by a plateau. These observations suggest that while minimizing the maximum of simultaneously active replicative proteins may be a factor that determines the replication kinetics, there must be a stronger limiting factor at the beginning of S phase to suppress the fast initial increase. A plausible hypothesis is that the copy number of some replicative proteins is small in the beginning of S phase but gradually increases with nuclear import [139]. This would lead to suppression of fast replication kinetics at the early stage of S phase. In conclusion, the optimization method presented here connects the replication process with an objective function that relates to evolutionary selection pressure and allows one to explore the limiting factors of replication.

6.5 The lattice-genome model: from random to periodic licensing

Until now, we have assumed a spatially random distribution of potential origins. In this section, we explore the implications of spatial ordering among the potential origins on the end-time distribution. We have two motivations. First, an "obvious" method for obtaining a narrow end-time distribution is to space the potential origins periodically and initiate them all at once. However, such an arrangement would not be robust, as the failure of just one origin to initiate would double the replication time. Still, the situation is less clear when initiations are spread out in time, as the role of spatial regularity in controlling inter-origin spacing is blurred by the temporal randomness.

Our second motivation is that there is experimental evidence that origins are not positioned completely at random. A completely random positioning implies that the distribution of gaps between potential origins is exponential, resulting in many small inter-potential-origin spacings. However, in an experiment of plasmid replication in *Xenopus* egg extracts, Lucas *et al.* found no inter-origin gap smaller than 2 kb [147]. In a previous analysis, Jun and Bechhoefer also observed that, assuming random licensing, one expects more inter-origin gaps less than 8 kb than were observed and fewer between 8–16 kb [34]. Moreover, experiments have suggested a qualitative tendency for origins to fire in groups, or clusters [138]. These findings collectively imply that there is some spatial regularity in the *Xenopus* system, perhaps through a "lateral inhibition" of licensing potential origins too closely together. Our goal is to find an "ordering threshold," at which point the resulting end-time distribution starts to deviate from the random-licensing case.

To investigate spatial ordering, we change the continuous genome to a "lattice genome" with variable lattice spacing d_l. Potential origins can be licensed only on the lattice sites. For $d_l \to 0$, the lattice genome becomes continuous, and the model recovers the random-licensing case. As d_l increases, the lattice genome has fewer available sites for licensing potential origins, and the fraction of licensed sites increases. In this scenario, the spacings between initiated origins take on discrete values—multiples of d_l. One can imagine that a further increase in d_l would eventually lead to a critical d_l, where every lattice site would have a potential origin. This scenario corresponds to an array of periodically licensed origins, which leads to a periodic array of initiated origins with spacing d_l. Thus, by increasing a single parameter d_l, we can continuously interpolate from complete randomness to perfect periodicity.

In order to compare regularized licensing to random licensing, we impose that while the potential origins may be distributed along the genome differently, the total initiation probability across the genome is conserved. We then write

$$I(x,t) = d_l\, I(t) \sum_{n=0}^{L/d_l} \delta(x - n d_l)\ , \tag{6.9}$$

where x is the position along the genome. Equation 6.9 shows that as the number of lattice sites L/d_l is reduced via an increase in d_l, the initiation probability for each site is enhanced, resulting in more efficient potential origins[†]. This implies a tradeoff between the "quantity"

[†] This concept is also captured by Eq. 2.13, Sec. 3.3.7, and the multiple-initiator model (MIM) in Chapter 4. Since this work was published before the study in Chapter 3 and 4, we did not try to explicate the connections in this chapter.

and "efficiency" of potential origins.

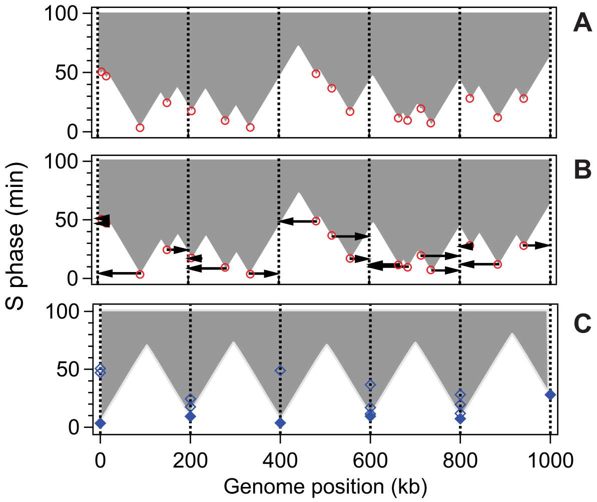


Figure 6.6: Schematic diagram of licensing on a lattice genome. **A**. A realization of replication using random-licensing ($d_l = 0$ case). The grey (white) area represents replicated (unreplicated) domains. Circles denote initiations. **B**. Origins are forced to their nearest lattice sites (marked by vertical lines at multiples of $d_l = 200$ kb), while initiation times remain the same. **C**. The result of the shift in origin positions. Open markers represent "phantom origins" that do not contribute to the replication; filled markers denote the actual origins. Alternatively, a filled marker can be viewed as the origin that initiated in a cluster of potential origins. Going from $d_l = 0$ kb in **A** to 200 kb in **C**, the average initiation time decreased from about 22 min to about 10 min.

Figure 6.6 shows how Eq. 6.9 connects random licensing to ordered licensing and illustrates this tradeoff. A realization of random licensing is shown in Fig. 6.6A. Since Eq. 6.9 modifies only the spatial distribution of origins relative to our previous $I(t)$, the effect of going from a continuous genome to a lattice genome is equivalent to shifting the randomly licensed origins to their nearest lattice sites while preserving their initiation times (Fig. 6.6B). In so doing, we obtained Fig. 6.6C, which shows multiple initiations on a lattice site. Since re-initiation is forbidden in normal replication, on each site only the earliest initiation contributes to the replication. The later initiations are "phantom origins" that illustrate how ordering reduces the number of initiations but enhances the efficiency of potential origin sites. The increase in efficiency is indicated by the decrease in the average initiation times between the two scenarios.

A perhaps more interesting and biologically relevant interpretation of Fig. 6.6 is that when potential origins cluster together, the one that initiates earliest can passively replicate the nearby potential origins. In other words, clustering can, in effect, reduce the effective number of potential origins but increase their efficiency. Thus, the increasing spatial order of potential origins from Fig. 6.6A to 6.6C can be interpreted either as having fewer but more efficient actual potential origins or as indicating clustering.

Having outlined the rules for licensing, we now introduce two quantities, "periodicity" P and d_{inter}, that will be useful in later discussions of how d_l alters the end-time distribution. We first look at $\rho_i(s)$, the distribution of the spacing between initiated origins, where s is the inter-origin spacing. Figure 6.7A shows two $\rho_i(s)$: the continuous one corresponds to random licensing, while the discrete one corresponds to setting d_l to 2 kb. The two distributions are different because of the discretization effect of the lattice genome: origins can have separations that are only multiples of d_l. As d_l increases, one expects a dominant spacing to appear in the system. We characterize this ordering effect by defining the periodicity P as the probability at the mode of the discrete inter-origin-spacing distribution. As an example, the $d_l = 2$ kb distribution shown in Fig. 6.7A has $P = 0.23$, indicating that 23% of the nearest neighbour origin pairs are equally spaced. In the fully periodic case, the probability at the mode is 1; all the spacings have the same value; and the system is 100% periodic ($P = 1$). For $d_l \to 0$, P should be interpreted as the mode of $\rho_i(s)$ times a vanishingly small Δs ($\sim d_l$). Thus, $P \to 0$ in the small Δs limit, as there will be no inter-origin spacings sharing the same size.

In interpolating from random licensing to periodic licensing, one expects that the average inter-origin spacing d_{avg} to change from being d_l-independent to being linearly dependent on d_l. Indeed, from Fig. 6.7B, which shows d_{avg} as a function of d_l, we can label two asymptotes and identify two regimes. We first introduce d_{inter} to be the average inter-origin spacing of the $d_l = 0$ kb case. For $d_l \to 0$, d_{avg} asymptotically approaches d_{inter}. In contrast, for large d_l (when all lattice sites are occupied), d_{avg} approaches the asymptote $d_{avg} = d_l$. The intersection of the two asymptotes is precisely at $d_l = d_{inter}$. We therefore identify two regimes, with Regime I being $d_l \leq d_{inter}$ and Regime II being $d_l > d_{inter}$. Physically, the weak d_l dependence in Regime I suggests that the system is still spatially random, whereas the asymptotically linear behaviour in Regime II indicates that the system is becoming periodic.

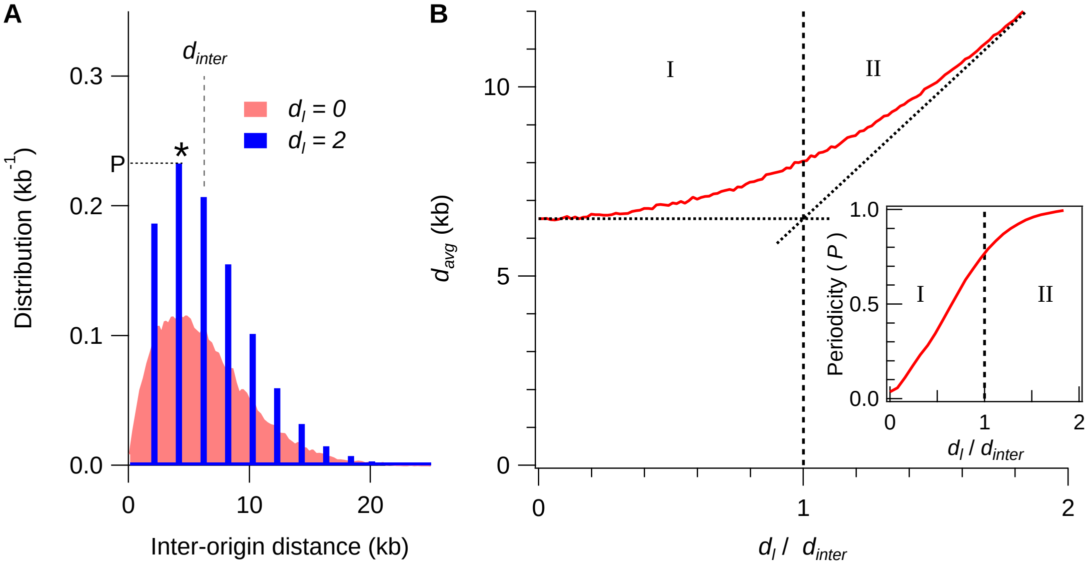


Figure 6.7: Properties of the lattice-genome model. **A**. The distribution of spacings between initiated origins, $\rho_i(s)$, for the $d_l = 0$ and 2 kb cases (2 kb is chosen to mimic the minimal spacing between origins reported in [147]). The initiation rate and fork velocity are those discussed in Sec. 6.3.2. The mean of the continuous distribution ($d_l = 0$ kb case) is marked d_{inter} and is $\approx$ 6.5 kb. The mode of the discrete distribution ($d_l = 2$ kb case) is marked by " $\star$ ". The probability P at the mode (0.23 in this case) is defined to be the periodicity, a measure of ordering in the system. **B**. Average inter-origin spacing d_{avg} as a function of d_l. There is a gradual transition from Regime I to Regime II. In Regime I ($d_l \leq d_{inter}$), d_{avg} is asymptotically independent of d_l for $d_l \rightarrow 0$. In Regime II ($d_l > d_{inter}$), d_{avg} is asymptotically linearly proportional to d_l. Inset shows the periodicity P as a function of d_l.

The length scale d_{inter} encodes the two factors that determine the distribution of inter-origin spacings. The first factor is the passive replication of closely positioned potential origins, which suppresses the likelihood of having small inter-origin spacings. The second factor is based on the low probability of randomly licensing two far-away origins, which reduces the probability of having large inter-origin gaps. Both of these effects can be seen in Fig. 6.7A.

When d_l exceeds d_{inter}, the typical spacing between potential origins ($\sim d_l$) exceeds the typical range of passive replication and approaches the typical largest spacing of the random-licensing case. This means that potential origins are not likely to be passively

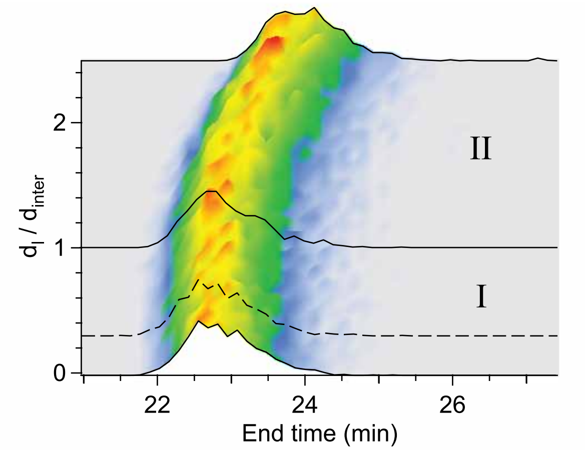


Figure 6.8: The end-time distribution as a function of spatial ordering defined by d_l. Each horizontal profile is an end-time distribution. In Regime I, the end-time distribution does not change appreciably; in Regime II, the mode shifts to the right. The ordering threshold is at $d_l = d_{inter} \approx 6.5$ kb. The dashed line shows the $d_l = 2$ kb end-time distribution, which corresponds to the lateral inhibition ordering observed experimentally [147].

replicated or positioned farther than d_l apart (note that the next smallest spacing $2d_l$ is quite large). The inset in Fig. 6.7B, which shows the periodicity P as a function of d_l, strengthens this notion that for $d_l > d_{inter}$, the system enters a nearly periodic regime where P has saturated.

Our main result is Fig. 6.8, which shows how the end-time distribution changes with increasing d_l. The initiation rate used in the simulation is the power-law approximation of the $I_{vivo}(t)$ mentioned in Sec. 6.2, transformed using Eq. 6.9. The fork velocity and failure rate used are as extracted in Sec. 6.3. There are again two distinct regimes separated by the ordering threshold $d_{inter} \approx 6.5$ kb. Below the threshold (Regime I), the end-time distribution is nearly independent of d_l. Above the threshold (Regime II), the mode shifts to the right. The width is unaltered.

To understand the changes in going from Regime I to Regime II, we note that in Eq. 2.19, t^* depends on the number of initiations N_o. On average, N_o is unaffected when the number of lattice sites available is in excess ($\frac{N_o}{L/d_l} > 1$). This means that t^* starts to change only when $d_l = L/N_o$ which is precisely d_{inter}. In Regime II, the minimum time to replicate the smallest gap between potential origins, d_l/v, becomes significant compared to the temporal randomness resulting from stochastic initiation. In effect, $t^* \approx d_l/v + t_{avg}$, where t_{avg} is the average initiation time. We tested numerically that the mean and standard deviation of the initiation times both decrease sigmoidally, for $d_l/d_{inter} > 3$. Thus, for the range of d_l shown in Fig. 6.8, one expects $t^* \propto d_l$ in Regime II, while the width should be unaltered.

In *Xenopus* embryos, the inhibition zone observed in plasmid replication corresponds to $d_l \approx 2$ kb (dashed line in Fig. 6.8) [147]. The value is well below the ordering threshold of $d_{inter} \approx 6.5$ kb, suggesting that the experimentally observed spatial ordering plays a minor role in solving the random-completion problem in embryonic replication.

In a very recent work, Karschau *et al.* studied replication completion in terms of optimization and spatial ordering as well [157]. They propose that the DNA replication process evolves to minimize the replication-completion time, as faster completion would allow faster development and provide evolutionary advantage. Their main result is that under a scenario where the efficiency of each potential origin is below a critical threshold, forming clusters of potential origins is the strategy to minimize the replication-completion time. They then argued qualitatively that the clustering scenario matches experimental observations. One issue with this proposal is that there is no clear mechanism that enforces the formation of the clusters described in the paper. In addition, the experimental observations, while qualitatively consistent with clustering, are also well captured by the random scenario (see Fig. 3C in [157]). In summary, stronger experimental evidence is needed to support a clustering scenario.

6.6 Conclusion

In this chapter, we have extended the stochastic nucleation-and-growth model of DNA replication to describe not only the kinetics of the bulk of replication but also the statistics of replication quantities at the end of replication. Using the model, we have quantitatively addressed a generalized version of the random-completion problem, which asks how stochastic licensing and initiation lead to the tight control of replication end times observed in systems such as *Xenopus* embryos. In particular, we applied our model to investigate and compare the two solutions proposed by biologists—the regular-spacing model (RSM) and the origin-redundancy model (ORM).

First, we found that the ORM, which utilizes purely random licensing, can still accurately control the replication time. With this approach, the fluctuation of the end times is suppressed by licensing a large excess of potential origins, while the typical end time is adjusted by increasing the initiation rate toward late S phase. Then, we analyzed the effect of spatial ordering in the RSM using a lattice genome. Our results show that 1) incorporating regularity leads to a tradeoff: the large number of potential origins in the ORM is effectively replaced by fewer but more efficient origins in the RSM and that 2) under the condition that the initiation rate across the genome is preserved, the two models produce the same end-time distribution until an ordering threshold is reached. We show that the experimentally observed ordering effect of lateral inhibition in *Xenopus* is well below the ordering threshold.

These results are particularly enlightening when considering clustering as a mechanism that transforms the ORM into the RSM. As discussed in Sec. 6.5, clustering spontaneously leads to a tradeoff between quantity and efficiency of potential origins while satisfying the condition of preserving the initiation rate. Thus, the intrinsic reason that the RSM and the ORM produce the same end-time distribution is not the spatial distribution of potential origins but the high density of potential origins. We argue that the key factors in resolving the random-completion problem, at least in the *Xenopus* case, are the licensing of a large excess of potential origins and an increasing initiation rate—and not an ordered spatial distribution of origins. To say it in a different way, the analysis in Sec. 6.5 implies that the end-time distribution due to a random ordering of potential origins would be unaltered if those same potential origins were positioned more regularly (e.g., in ordered clusters), as long as the regularity is below a threshold which exceeds the experimentally observed amounts in frog embryos.

We have also found the optimal $I(t)$ that minimizes the maximum number of simultaneously active forks. Similar to the observed *in-vitro* initiation rate, it increases throughout S phase except for the end. Further pursuit of the optimization problem with more detailed model may reveal the rate-limiting factors in replication, which have not been identified to date. An open issue not addressed by our model is the observation that there is a weak correlation in the initiations of neighbouring origins [138] via chromatin structure [62], fork progression [158, 123, 25], or other unknown mechanisms. We do not expect that correlations will modify the scenario we have presented here significantly, as the most significant effect of correlations, an increase in spatial ordering, would not be important even at exclusion-zone sizes that are much larger than observed (e.g., 10 kb).

Have we fully solved the random-completion problem? Recall that there is on average one ORC every 8 kb of DNA, and each ORC loads 20–40 MCMs. Counting a pair of MCMs as a potential origin, there is, on average, roughly two potential origins every kb. According to our results, if these MCMs were to move along the DNA and spread out, the cells would not need a mechanism that regularizes the spacing between ORC or MCM. Experiments showed that MCM can slide along free DNA molecules *in vitro* [9, 10]; however, little is known about the mobility of MCMs *in vivo*, where the DNA form higher-order structures in order to fit inside the nucleus. On the other hand, if loaded MCMs were constrained to be near an ORC, the cells would need either a mechanism that regularizes the spacing between the ORCs or a mechanism that allows an ORC to visit multiple loci. Since the distribution of MCM *in vivo* is unknown, our analysis cannot rule out the regular-spacing model. Nevertheless, we have shown that the two models (regular-spacing model and origin-redundancy model), though mechanistically very different, result in essentially the same end-time distribution.

Connecting the solution of the random-completion problem to our results in Chapter 3, we note that the replication in budding yeast also uses redundant potential origins that have increasing initiation rates. Although budding yeast, unlike embryonic cells, has many cell cycle checkpoints, there is no direct evidence for a checkpoint that detects incomplete duplication. To our knowledge, the known S/M checkpoints all respond to replication stress such as DNA breakages and fork stalls (e.g., [159, 160]). If there were really no S/M checkpoint for normal replication, would budding yeast suffer from the fluctuation in replication-completion times? We simulated the replication process using the SM parameters and found that the mode and width of the end-time distribution are roughly 75 and 10 min, respectively. The cell cycle of the experimentally probed culture is roughly 200 min [52], and S and G2 phase together usually constitute $\sim$50% of the cell cycle [161]. Thus, under normal circumstances (without replication stress), duplication essentially always finishes before mitosis. Putting the results for frog embryos and budding yeast in a broader perspective, we propose that instead of forming regulatory feedback mehcanisms to ensure replication completion, eukaryotic cells complete genome duplication in a feedforward manner by adopting the proper initiation rates.

Chapter 7

Conclusion

Replication kinetics in eukaryotes is difficult to understand without a quantitative framework because replication starts at many sites across the genome and throughout S phase in a stochastic manner. In this thesis, we have first developed a general mathematical framework based on the stochastic nucleation-and-growth theory introduced by Kolmogorov, Johnson, Mehl, and Avrami. Our theory includes probabilistic licensing and initiation and the effect of passive replication. The probabilistic nature of the model allows analysis of a broad range of kinetics—from deterministic to random origin positioning and timing—and is thus general enough to describe eukaryotic replication. Accounting for passive replication allows one to characterize the potential efficiency of origins and explain how the replication process is robust against replication stress. We also derived the distribution of replication-completion times, a result that allows us to understand how genome duplication connects with the replication program.

A major motivation for developing the theory is to apply it to experiments. Because our theory is analytic, it allows efficient information extraction from genome-wide experiments. In this thesis, we have applied our theory to two experimental techniques that probe the genome-wide progress of replication, namely time-course microarray (Chapter 3) and FACS-microarray (Chapter 5). We believe that the methods presented in this thesis, particularly the analysis for FACS-microarray, can be used to extract quantitative information about replication kinetics in many organisms.

We specifically built models for two organisms: budding yeast and frog embryos. In the case of budding yeast, we fit the Sigmoid Model (SM) to a recent time-course dataset and extracted a striking feature—that the average firing time of an origin correlates strongly with its timing precision. This finding has at least three major implications: 1) the initiation timing in budding yeast is not deterministic; 2) earlier-firing origins are more efficient; and 3) there is a global mechanism that underlies origin firing in addition to local variations. Based on this result, we proposed the Multiple-Initiator Model (MIM) and showed that reproducible replication patterns need not result from time-measuring activators but can be a consequence of random and identical activators and initiators. The details of the model suggest a specific molecular mechanism for replication timing, in which the number of minichromosome maintenance (MCM) complexes loaded at an origin and the chromatin structure around the origin together regulate the origin firing time.

For the case of frog embryos, we showed that the random-completion problem, which asks how the replication-completion time can be tightly controlled in the presence of random licensing and initiation, is solved by 1) licensing more potential origins than needed and 2) adopting an initiation rate that increases throughout most of S phase. Interestingly, we found that the initiation rate extracted from experiment not only exhibits these two properties but also nearly optimizes replication resources. Lastly, compared to the two properties mentioned, we showed that spatial regularity is not an important factor in controlling the genome-duplication time in frog embryos.

As mentioned in Sec. 1.1, the replication kinetics in budding yeast and frog embryos are often thought to be at the two ends of a “deterministic-to-random” spectrum. Licensing in budding yeast is associated with particular sequences, while licensing in frog embryos is sequence independent. Interestingly, although the licensing properties of the two organisms differ drastically, the control of replication timing is very similar. In budding yeast, we showed that the average and precision of origin timing are related to the number of initiators loaded onto the origins. In frog embryos, we showed that the average and precision of replication-completion time are related to the number of potential origins licensed. In other words, in both organisms, replication timing is controlled by the number of initiating elements. Also, the initiation rates in both organisms increase throughout most of S phase—a strategy that safeguards against fork stalls and large unreplicated gaps. These two mechanisms—loading excessive initiators and adopting an increasing initiation rate—offer a simple and plausible explanation to robust timing control that contrasts with the popular but mechanistically elusive view that replication timing is controlled carefully by mechanisms that measure time and origin spacing.

7.1 Future directions

7.1.1 Universal features of eukaryotic replication

Goldar *et al.* showed evidence for a universal replication program across many species [38]; however, the analysis in [38] was deterministic and averaged over the genome. We propose that using the methods presented in this thesis can lead to more reliable estimates of the replication program. In particular, one could test whether the correlation between initiation-timing average and precision is a universal feature across many species.

In order to do this, we first note that one can speed up the SM and MIM fit by exploiting the parameter structure. As mentioned in Sec. 3.2 and Sec. 4.2.1, both models consist of a handful of global parameters and many local parameters associated with the origins. In the present algorithm, all parameters, including the local origin parameters, are fit globally. However, since an origin cannot affect the replication kinetics far away from it, performing a global evaluation of χ^2 is a waste of computational power. We thus suggest that the fit can be separated into two routines: one performs a global search for the global parameters, while the other performs a local search for the origin parameters and sweeps through the genome. We note that the local search cannot be done origin by origin because the origins are weakly coupled together via passive replication. Nevertheless, if the local search were properly programmed, we expect that the fit would speed up by orders of magnitude so that it could be applied to larger genomes of multicellular organisms.

The SM and MIM, in their present forms, need prior information on the number of origins. In Chapter 5, we presented a more general method that relaxes this requirement. The method is based on constrained optimization and assumes only generic structures such as bi-directional fork propagation and smooth initiation rates. We demonstrated that the method can extract information about the replication program from simulated FACS-microarray data and pointed out specific ways to improve the method. An implementation of the improved method for large datasets will allow one to extract replication information from real FACS-microarray experiments, which are much more accessible than time-course microarray. At present, there are at least 100 FACS-microarray datasets on eight species

[162, 14] available already! We believe that applying the improved and extended method to these datasets and the many more to come will reveal a more detailed kinetic picture of eukaryotic replication in general.

With respect to testing the correlation between initiation timing and precision, our analysis showed that the simplest FACS-microarray data, which provide only two replication fraction profiles (one averaged over the genome and the other over most of S phase), do not contain enough information (see Sec. 5.4.1). However, more elaborated FACS-microarray experiments can provide multiple replication fraction profiles that are averaged over narrower sections of S phase. In fact, such experiments have already been done (e.g., [163]). We expect that the improved and extended method can be applied to these FACS-microarray datasets to provide information about the correlation. With this information, one can start to test the hypothesis that the correlation mentioned above is a universal feature of eukaryotic replication.

7.1.2 Incorporating replicative machinery

In this thesis, we focused on analyzing experiments that probe the fraction of replication. We propose that extending the analysis to simultaneously incorporate other replicative machinery can yield even more details. For example, by incorporating the time-course profiles of fork density in [48] with the replication profiles in [57], one can better estimate the spatiotemporal variations in fork velocity and hence also the replication program. (The theory based on fork density is developed in [17].) If one further incorporates the time-course profiles of MCM occupancy, one can start to address whether the MCM complexes are pushed along or pushed off the DNA by replication forks. If the MCM are pushed along, this can lead to a correlation between fork progression and origin initiation. In [25], Ma *et al.* proposed such a correlation for budding yeast. In any case, the answer to this question can further refine our understanding of the replication kinetics.

7.1.3 Incorporating three-dimensional structure

Until now, most mathematical and simulation models for DNA replication kinetics (see [8] for a review) have ignored the three-dimensional structure of the chromosomes and the crowded nuclear environment. However, recent evidence shows strong correlations between the origin function and nucleosome positioning [47] and between replication timing and chromatin structure [164]. In fact, the latter correlation is one of the strongest correlations in genomics [162]. The chromatin structure is characterized by a newly developed technique called Hi-C that measures the contact frequencies among chromosomes [120]. Using the Hi-C techniques, Lieberman-Aiden *et al.* confirmed the well-established observation that chromatin in humans is compartmentalized with two distinct structures: a euchromatin structure that is highly accessible and and a heterochromatin structure that is more compact [120]. Using the same technique, Duan *et al.* proposed a static three-dimensional model of the budding yeast genome inside the nucleus [165]. A physical theory that incorporates such three-dimensional information, perhaps via polymer physics, would be an exciting future direction. As an example, one can model replicative proteins using strictly diffusive particles that have certain binding and unbinding rates and model chromosomes using an interacting bead-on-string model where the interactions among the beads determine the chromatin accessibility. Such models would allow comparison with many Hi-C and microscopy studies that relate chromatin structure and nuclear positioning to replication timing [70, 71, 72, 164, 166].

Furthermore, in Chapter 4, we proposed that the timing of origin initiation is determined by the number of initiators loaded onto the origins and the chromatin structure around the origin. Although we showed that, in budding yeast, there is significant correlation between the timing of initiation and the number of origin-bound MCM (a pair of which is a biologically plausible initiator), we cannot explain all the variations in initiation timing by the MCM occupancy. A detailed investigation of the three-dimensional chromatin structures within the nucleus during S phase might reveal the missing factors.

7.1.4 Abnormal replication

As mentioned in Sec. 1.2, we hope that modelling DNA replication kinetics can contribute to understanding and eventually treating cancer. This would require a sound understanding of replication under different conditions. Experiments have, in fact, probed the replication fraction of budding yeast in many mutants [53]. We believe that modelling these datasets with the methods presented in this thesis can advance and solidify our understanding of normal replication.

To model replication in cancer cells, one needs to extend the theory presented here to incorporate abnormal kinetic elements such as fork stalls and re-replication. Previous work that modelled the effect of fork stalls has already shown that cancerous cells have higher stall densities than do normal cells [46]. Also, since cancer is a developmental disease, one might need a theory for mitosis and checkpoint activation that allows the effect of abnormal replication to be propagated or suppressed in order to address the emergence of cancer.

7.2 A final remark

Joel Cohen has stated that “Mathematics is biology’s next microscope, only better” [167]. Just as a microscope allows one to see things too small to see with the naked eye, mathematics allows one to obtain information that would be impossible to infer using purely qualitative arguments. The analyses presented in this thesis illustrate this concept. In Chapter 3, by applying our mathematical model to genome-wide replication experiment probing budding yeast, we revealed a correlation between the average and the precision of initiation timing. Without such quantitative analysis [74, 42, 23], the correlation would have remained hidden in the data, as it had for the past decade since the publication of the first genome-wide replication data in 2001 [20]. Further modelling of the correlation then led to the first transparent molecular mechanism for controlling initiation timing. In a video entitled “A vision for quantitative biology” in i-Bio-Magazine, Rob Phillips argues that mathematics can sharpen the kind of questions we ask about biology [168]. We saw such sharpened understanding emerge from our analysis of the random-completion problem in Chapter 6. There, we found that since nearly any initiation rate can satisfy the replication timing constraints, the problem is not only to complete the duplication on time but to complete it while optimizing replicative resources. Through these examples, I hope to have convinced you, the reader, that mathematical modelling is essential to the understanding of the important, complex, and—when seen in the right light—rather beautiful dynamical processes that underlie life itself.

Bibliography

[1] Hensey, C. & Gautier, J. A developmental timer that regulates apoptosis at the onset of gastrulation. *Mech. Dev.* **69**:183–195 (1997).

[2] Micco, R. D., Fumagalli, M., Cicalese, A., Piccinin, S., Gasparini, P., Luise, C., Schurra, C., Garre', M., Nuciforo, P. G., Bensimon, A., Maestro, R., Pelicci, P. G. & d'Adda di Fagagna, F. Oncogene-induced senescence is a DNA damage response triggered by DNA hyper-replication. *Nature* **444**:638–642 (2006).

[3] Sancar, A., Lindsey-Boltz, L. A., Ünsal Kaçmaz, K. & Linn, S. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu. Rev. Biochem.* **73**:39–85 (2004).

[4] Arias, E. E. & Walter, J. C. Strength in numbers: preventing rereplication via multiple mechanisms in eukaryotic cells. *Genes Dev.* **21**:497–518 (2007).

[5] Sclafani, R. A. & Holzen, T. M. Cell cycle regulation of DNA replication. *Annu. Rev. Genet.* **41**:237–280 (2007).

[6] Lodish, H., Berk, A., Kaiser, C. A., Krieger, M., Scott, M. P., Bretscher, A., Ploegh, H. & Matsudairu, P. *Molecular Cell Biology* (W.H. Freeman and Company, New York, NY, 2008), 6th ed.

[7] Herrick, J. & Bensimon, A. Global regulation of genome duplication in eukaryotes: an overview from the epifluorescence microscope. *Chromosoma* **117**:243–260 (2008).

[8] Hyrien, O. & Goldar, A. Mathematical modelling of eukaryotic DNA replication. *Chromosome Res.* **18**:147–161 (2010).

[9] Remus, D., Beuron, F., Tolun, G., Griffith, J. D., Morris, E. P. & Diffley, J. F. Concerted loading of Mcm2–7 double hexamers around DNA during DNA replication origin licensing. *Cell* **139**:719–730 (2009).

[10] Evrin, C., Clarke, P., Zech, J., Lurz, R., Sun, J., Uhle, S., Li, H., Stillman, B. & Speck, C. A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. *Proc. Natl. Acad. Sci.* **106**:20240–20245 (2009).

[11] Waga, S. & Stillman, B. The DNA replication fork in eukaryotic cells. *Annu. Rev. Biochem.* **67**:721–751 (1998).

[12] Kunkel, T. A. & Burgers, P. M. Dividing the workload at a eukaryotic replication fork. *Trends Cell Biol.* **18**:521–527 (2008).

[13] Dai, J., Chuang, R.-Y. & Kelly, T. J. DNA replication origins in the *Schizosaccharomyces pombe* genome. *Proc. Natl. Acad. Sci.* **102**:337–342 (2005).

[14] MacAlpine, D. M., Rodríguez, H. K. & Bell, S. P. Coordination of replication and transcription along a *Drosophila* chromosome. *Genes Dev.* **18**:3094–3105 (2004).

[15] MacAlpine, H., Gordan, R., Powell, S., Hartemink, A. & MacAlpine, D. *Drosophila* ORC localizes to open chromatin and marks sites of cohesin loading. *Genome Res.* **20**:201–211 (2009).

[16] Woodfine, K., Fiegler, H., Beare, D. M., Collins, J. E., McCann, O. T., Young, B. D., Debernardi, S., Mott, R., Dunham, I. & Carter, N. P. Replication timing of the human genome. *Hum. Mol. Gen.* **13**:191–202 (2004).

[17] Gauthier, M. G., Norio, P. & Bechhoefer, J. Modeling inhomogeneous DNA replication kinetics. *PLoS ONE* **7**:e32053 (2012).

[18] 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.* **12**:4511–4520 (1993).

[19] Labit, H., Perewoska, I., Germe, T., Hyrien, O. & Marheineke, K. DNA replication timing is deterministic at the level of chromosomal domains but stochastic at the level of replicons in *Xenopus* egg extracts. *Nucleic Acids Res* **36**:5623–5634 (2008).

[20] Raghuraman, M. K., Winzeler, E. A., Collingwood, D., Hunt, S., Wodicka, L., Conway, A., Lockhart, D. J., Davis, R. W., Brewer, B. J. & Fangman, W. L. Replication dynamics of the yeast genome. *Science* **294**:115–121 (2001).

[21] Patel, P. K., Arcangioli, B., Baker, S. P., Bensimon, A. & Rhind, N. DNA replication origins fire stochastically in fission yeast. *Mol. Biol. Cell* **17**:308–316 (2006).

[22] Czajkowsky, D. M., Liu, J., Hamlin, J. L. & Shao, Z. DNA combing reveals intrinsic temporal disorder in the replication of yeast chromosome VI. *Mol. Biol. Cell.* **375**:12–19 (2008).

[23] Brümmer, A., Salazar, C., Zinzalla, V., Alberghina, L. & Höfer, T. Mathematical modelling of DNA replication reveals a trade-off between coherence of origin activation and robustness against rereplication. *PLoS Comp. Biol.* **6**:e1000783 (2010).

[24] Bertuzzi, A., Gandolfi, G., Germani, A. & Vitelli, R. A general expression for sequential DNA-fluorescence histograms. *J. Theor. Biol.* **102**:55–67 (1983).

[25] Ma, E., Hyrien, O. & Goldar, A. Do replication forks control late origin firing in *Saccharomyces cerevisiae*? *Nucl. Acids Res.* **40**:2010–2019 (2012).

[26] Cowan, R. *Stochastic processes: modelling and simulation*, chap. 4. Stochastic models for DNA replication, pp. 137–166 (Elsevier, Boston, MA, 2003).

[27] Kolmogorov, A. N. Static theory of metals crystallization. *Bull. Acad. Sc. USSR, Phys. Ser.* **1**:335 (1937).

[28] Johnson, W. A. & Mehl, F. L. Reaction kinetics in processes of nucleation and growth. *Trans. AIME* **135**:416 (1939).

[29] Avrami, M. Kinetics of phase change. I General theory. *J. Chem. Phys.* **7**:1103 (1939).

[30] Sekimoto, K. Kinetics of magnetization switching in a 1-D system-size distribution of unswitched domains. *Physica A* **125**:261–269 (1984).

[31] Sekimoto, K. Evolution of the domain structure during the nucleation-and-growth process with non-conserved order parameter. *Physica A* **135**:328–346 (1986).

[32] Ben-Naim, E. & Krapivsky, P. L. Nucleation and growth in one dimension. *Phys. Rev. E* **54**:3562–3568 (1996).

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

[34] Jun, S. & Bechhoefer, J. Nucleation and growth in one dimension. II. Application to DNA replication kinetics. *Phys. Rev. E* **71**:011909 (2005).

[35] Yang, S. C.-H., Gauthier, M. G. & Bechhoefer, J. *DNA Replication: Methods and Protocols*, vol. 521 of *Methods in Molecular Biology*, chap. Computational methods to study kinetics of DNA replication, pp. 555–574 (Humana Press, c/o Springer Science + Business Media, New York, NY, 2009).

[36] Herrick, J. & Bensimon, A. Single molecule analysis of DNA replication. *Biochimie* **81**:859–871 (1999).

[37] Gauthier, M. G. & Bechhoefer, J. Control of DNA replication by anomalous reaction-diffusion kinetics. *Phys. Rev. Lett.* **102**:158104 (2009).

[38] Goldar, A., Marsolier-Kergoat, M.-C. & Hyrien, O. Universal temporal profile of replication origin activation in eukaryotes. *PLoS ONE* **4**:e5899 (2009).

[39] Eshaghi, M., Karuturi, R. K. M., Li, J., Chu, Z., Liu, E. T. & Liu, J. Global profiling of DNA replication timing and efficiency reveals that efficient replication/firing occurs late during S-phase in *S. pombe*. *PLoS ONE* **2**:e722 (2007).

[40] Lygeros, J., Koutroumpas, K., Dimopoulos, S., Legouras, I., Kouretas, P., Heichinger, C., Nurse, P. & Lygerou, Z. Stochastic hybrid modeling of DNA replication across a complete genome. *Proc. Natl. Acad. Sci.* **105**:12295–12300 (2008).

[41] Blow, J. J. & Ge, X. Q. A model for DNA replication showing how dormant origins safeguard against replication fork failure. *EMBO Rep.* **10**:406–412 (2009).

[42] de Moura, A. P. S., Retkute, R., Hawkins, M. & Nieduszynski, C. A. Mathematical modelling of whole chromosome replication. *Nucl. Acids Res.* **38**:5623–5633 (2010).

[43] Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability—an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.* **11**:220–228 (2010).

[44] Halazonetis, T. D., Gorgoulis, V. G. & Bartek, J. An oncogene-induced DNA damage model for cancer development. *Science* **319**:1352–1355 (2008).

[45] Aguilera, A. & Gómez-González, B. Genome instability: a mechanistic view of its causes and consequences. *Nat. Rev. Genet.* **9**:204–217 (2008).

[46] Gauthier, M. G., Herrick, J. & Bechhoefer, J. Defects and DNA replication. *Phys. Rev. Lett.* **104**:218104 (2010).

[47] Eaton, M. L., Galani, K., Kang, S., Bell, S. P. & MacAlpine, D. M. Conserved nucleosome positioning defines replication origins. *Genes Dev.* **24**:748–753 (2010).

[48] Sekedat, M. D., Fenyö, D., Rogers, R. S., Tackett, A. J., Aitchison, J. D. & Chait, B. T. GINS motion reveals replication fork progression is remarkably uniform throughout the yeast genome. *Mol. Syst. Biol.* **6**:353 (2010).

[49] Futcher, B. Cell cycle synchronization. *Methods Cell Science* **21**:79–86 (1999).

[50] Kriegstein, H. J. & Hogness, D. S. Mechanism of DNA replication in *Drosophila* chromosomes: Structure of replication forks and evidence for bidirectionality. *Proc. Natl. Acad. Sci.* **71**:135–139 (1974).

[51] Huberman, J. A. & Riggs, A. D. Autoradiography of chromosomal DNA fibers from Chinese hamster cells. *Proc. Natl. Acad. Sci.* **55**:599–606 (1966).

[52] Brewer, B. J. & Fangman, W. L. The localization of replication origins on ARS plasmids in *S. cerevisiae*. *Cell* **51**:463–471 (1987).

[53] Koren, A., Soifer, I. & Barkai, N. MRC1-dependent scaling of the budding yeast DNA replication timing program. *Genome Res.* **20**:781–790 (2010).

[54] Herrick, J., Jun, S., Bechhoefer, J. & Bensimon, A. Kinetic model of DNA replication in eukaryotic organisms. *J. Mol. Biol.* **320**:741–750 (2002).

[55] Norio, P. & Schildkraut, C. L. Visualization of DNA replication on individual Epstein-Barr virus episomes. *Science* **294**:2361–2364 (2001).

[56] Heichinger, C., Penkett, C. J., Bähler, J. & Nurse, P. Genome-wide characterization of fission yeast DNA replication origins. *EMBO J.* **25**:5171–5179 (2006).

[57] McCune, H. J., Danielson, L. S., Alvino, G. M., Collingwood, D., Delrow, J. J., Fangman, W. L., Brewer, B. J. & Raghuraman, M. K. The temporal program of chromosome replication: genomewide replication in clb5Δ *Saccharomyces cerevisiae*. *Genetics* **180**:1833–1847 (2008).

[58] Desprat, R., Thierry-Mieg, D., Lailler, N., Lajugie, J., Schildkraut, C., Thierry-Mieg, J. & Bouhassira, E. E. Predictable dynamic program of timing of DNA replication in human cells. *Genome Res.* **19**:2288–2299 (2009).

[59] Shapiro, H. M. *Practical Flow Cytometry* (Wiley-Liss, New York, NY, 2003).

[60] Bensimon, A., Simon, A., Chiffaudel, A., Croquette, V., Heslot, F. & Bensimon, D. Alignment and sensitive detection of DNA by a moving interface. *Science* **265**:2096–2098 (1994).

[61] Herrick, J., Stanislawski, P., Hyrien, O. & Bensimon, A. Replication fork density increases during DNA synthesis in *X. laevis* egg extracts. *J. Mol. Biol.* **300**:1133–1142 (2000).

[62] Jun, S., Herrick, J., Bensimon, A. & Bechhoefer, J. Persistence length of chromatin determines origin spacing in *Xenopus* early-embryo DNA replication: Quantitative comparisons between theory and experiment. *Cell Cycle* **3**:223 (2004).

[63] Zhang, H. & Bechhoefer, J. Reconstructing DNA replication kinetics from small DNA fragments. *Phys. Rev. E* **73**:051903 (2006).

[64] Shendure, J. The beginning of the end for microarrays? *Nature Methods* **5**:585–587 (2008).

[65] van Oijen, A. M. & Loparo, J. J. Single-molecule studies of the replisome. *Annu. Rev. Biophys.* **39**:429–448 (2010).

[66] Lee, J.-B., Hite, R. K., Hamdan, S. M., Xie, X. S., Richardson, C. C. & van Oijen, A. M. DNA primase acts as a molecular brake in DNA replication. *Nature* **439**:621–624 (2006).

[67] Tanner, N. A., Loparo, J. J., Hamdan, S. M., Jergic, S., Dixon, N. E. & van Oijen, A. M. Real-time single-molecule observation of rolling-circle DNA replication. *Nucl. Acids Res.* **37**:e27 (2009).

[68] Malyavantham, K. S., Bhattacharya, S., Alonso, W. D., Acharya, R. & Berezney, R. Spatio-temporal dynamics of replication and transcription sites in the mammalian cell nucleus. *Chromosoma* **117**:553–567 (2008).

[69] Leonhardta, H., Rahna, H.-P., Weinzierlb, P., Sporberta, A., Cremerb, T., Zinkb, D. & Cardosoa, M. C. Dynamics of DNA replication factories in living cells. *J. Cell Biol.* **149**:271–280 (2000).

[70] Kitamura, E., Blow, J. J. & Tanaka, T. U. Live-cell imaging reveals replication of individual replicons in eukaryotic replication factories. *Cell* **125**:1297–1308 (2006).

[71] Zink, D. The temporal program of DNA replication: new insights into old questions. *Chromosoma* **115**:273 (2006).

[72] Baddeley, D., Chagin, V. O., Schermelleh, L., Martin, S., Pombo, A., Carlton, P. M., Gahl, A., Domaing, P., Birk, U., Leonhardt, H., Cremer, C. & Cardoso, M. C. Measurement of replication structures at the nanometer scale using super-resolution light microscopy. *Nucl. Acids Res.* **38**:e8 (2010).

[73] Yang, S. C.-H. & Bechhoefer, J. How *Xenopus laevis* embryos replicate reliably: investigating the random-completion problem. *Phys. Rev. E* **78**:041917 (2008).

[74] Yang, S. C.-H., Rhind, N. & Bechhoefer, J. Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. *Mol. Syst. Biol.* **6**:404 (2010).

[75] Guilbaud, G., Rappailles, A., Baker, A., Chen, C.-L., Arneodo, A., Goldar, A., d'Aubenton Carafa, Y., Thermes, C., Audit, B. & Hyrien, O. Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. *PLoS Comp. Biol.* **7**:e1002322 (2011).

[76] Baker, A. *Linking the DNA Strand Asymmetry to the Spatio-temporal Replication Program: From Theory to the Analysis of Genomic and Epigenetic Data*. Ph.D. thesis, Université de Lyon (2011).

[77] Branzei, D. & Foiani, M. The DNA damage response during DNA replication. *Curr. Opin. Cell Biol.* **17**:568–575 (2005).

[78] Danilowicz, C., Coljee, V. W., Bouzigues, C., Lubensky, D. K., Nelson, D. R. & Prentiss, M. DNA unzipped under a constant force exhibits multiple metastable intermediates. *Proc. Natl. Acad. Sci.* **100**:1694–1699 (2003).

[79] Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W. *Spikes: Exploring the Neural Code* (The MIT press, Cambridge, MA, 1999).

[80] Evans, M., Hastings, N. & Peacock, B. *Statistical Distributions* (Wiley-Liss, New York, NY, 2000), 3rd ed.

[81] Retkute, R., Nieduszynski, C. A. & de Moura, A. Dynamics of DNA replication in yeast. *Phys. Rev. Lett.* **107**:068103 (2011).

[82] Kotz, S. & Nadarajah, S. *Extreme Value Distributions: Theory and Applications* (Imperial College Press, River Edge, NJ, London, 2000).

[83] Gumbel, E. J. *Statistics of Extremes* (Columbia University Press, New York, NY, New York, 1958).

[84] Reiss, R.-D. & Thomas, M. *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields* (Birkhäuser Verlag, Basel, Basel, 2007), 3rd ed.

[85] Caldarelli, G., DiTolla, F. D. & Petri, A. Self-organization and annealed disorder in a fracturing process. *Phys. Rev. Lett.* **77**:2503 (1996).

[86] Moreira, A. A., Andrade, J. S. & NunesAmaral, L. A. Extremum statistics in scale-free network models. *Phys. Rev. Lett.* **89**:268703 (2002).

[87] Fréchet, M. Sur la loi de probabilité de l'écart maximum. *Ann. Soc. Polon. Math.* **6**:92–116 (1927).

[88] Fisher, R. A. & Tippett, L. H. C. Limiting forms of the frequency distribution of the largest or smallest number of a sample. *Proc. Cambridge Phil. Soc.* **24**:180–190 (1928).

[89] Tijms, H. *Understanding Probability: Chance Rules in Everyday Life* (Cambridge University Press, Cambridge, NY, Cambridge, 2004).

[90] Sethna, J. P. *Statistical Mechanics: Entropy, Order Parameters, and Complexity* (Oxford University Press, New York, NY, Oxford, 2006). The derivation of the Gumbel distribution appears in a web-supplement.

[91] MacKay, D. J. *Information Theory, Inference, and Learning Algorithms* (Cambridge University Press, Cambridge, UK, 2003).

[92] Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. *Numerical Recipes: The Art of Scientific Computing* (Cambridge University Press, Cambridge, UK, 2007), 3rd ed.

[93] WaveMetrics (2008). WaveMetrics Inc.

[94] Goren, A. & Cedar, H. Replicating by the clock. *Nat. Rev. Mol. Cell Biol.* **4**:25–32 (2003).

[95] Donaldson, A. D. Shaping time: chromatin structure and the DNA replication programme. *Trends Genet.* **21**:444–449 (2005).

[96] Alvino, G. M., Collingwood, D., Murphy, J. M., Delrow, J., Brewer, B. J. & Raghuraman, M. K. Replication in hydroxyurea: It's a matter of time. *Mol. Cell. Biol.* **27**:6396–6404 (2007).

[97] Nieduszynski, C. A., ichiro Hiraga, S., Ak, P., Benham1, C. J. & Donaldson, A. D. OriDB: a DNA replication origin database. *Nucleic Acids Res.* **35**:D40–D46 <`http://www.oridb.org`> (2007).

[98] Nasmyth, K. At the heart of the budding yeast cell cycle. *Trends Genet.* **12**:405–412 (1996).

[99] Hyrien, O., Marheineke, K. & Goldar, A. Paradoxes of eukaryotic DNA replication: MCM proteins and the random completion problem. *BioEssays* **25**:116–125 (2003).

[100] Rhind, N., Yang, S. C.-H. & Bechhoefer, J. Reconciling stochastic origin firing with defined replication timing. *Chromosome Res.* **18**:35–43 (2009).

[101] Ferguson, B. M., Brewer, B. J., Reynolds, A. E. & Fangman, W. L. A yeast origin of replication is activated late in S phase. *Cell* **65**:507–515 (1991).

[102] Blow, J. J., Ge, X. Q. & Jackson, D. A. How dormant origins promote complete genome replication. *Trends Biomed. Sci.* **36**:405–414 (2011).

[103] Meselson, M. & Stahl, F. W. The replication of DNA in *Escherichia coli*. *Proc. Natl. Acad. Sci.* **44**:671–682 (1958).

[104] Niemistö, A., Nykter, M., Aho, T., Jalovaara, H., Marjanen, K., Ahdesmäki, M., Ruusuvuori, P., Tiainen, M., Linne, M.-L., & Yli-Harja, O. Computational methods for estimation of cell cycle phase distribution of yeast cells. *EURASIP J. on Bioinfor Sys Biol* **2007**:46150 (2007).

[105] Orlando, D., Lin, C. Y., Bernard, A., Iversen, E. S., Hartemink, A. J. & Haase, S. B. A probabilistic model for cell cycle distributions in synchrony experiments. *Cell Cycle* **6**:478–488 (2007).

[106] Rivin, C. J. & Fangman, W. L. Replication fork rate and origin activation during the S phase of *Saccharomyces cerevisiae*. *J. Cell Biol.* **85**:108–115 (1980).

[107] Sivia, D. S. & Skilling, J. *Data Analysis: A Bayesian Tutorial* (Oxford University Press, New York, NY, 2006).

[108] Edwards, M. C., Tutter, A. V., Cvetic, C., Gilbert, C. H., Prokhorova, T. A. & Walter, J. C. MCM2–7 complexes bind chromatin in a distributed pattern surrounding the origin recognition complex in *Xenopus* egg extracts. *J. Biol. Chem.* **277**:33049–33057 (2002).

[109] Bowers, J., Randell, J., Chen, S. & Bell, S. ATP hydrolysis by ORC catalyzes reiterative Mcm2–7 assembly at a defined origin of replication. *Mol. Cell* **22**:967–978 (2004).

[110] Kerem, B. S., Goitein, R., Diamond, G., Cedar, H. & Marcus, M. Mapping of DNAase I sensitive regions on mitotic chromosomes. *Cell* **38**:493–499 (1984).

[111] Gómez-Llorente, Y., Fletcher, R. J., Chen, X. S., Carazo, J. M. & Martín, C. S. Polymorphism and double hexamer structure in the archaeal minichromosome maintenance (MCM) helicase from *Methanobacterium thermoautotrophicum*. *J. Biol. Chem.* **280**:40909–40915 (2005).

[112] Sharov, V., Kwong, K. Y., Frank, B., Chen, E., Hasseman, J., Gaspard, R., Yu, Y., Yang, I. & Quackenbush, J. The limits of log-ratios. *BMC Biotechnology* **4**:3 (2004).

[113] Xu, W., Aparicio, J. G., Aparicio, O. M. & Tavaré, S. Genome-wide mapping of ORC and Mcm2p binding sites on tiling arrays and identification of essential ARS consensus sequences in *S. cerevisiae*. *BMC Genomics* **7**:276 (2006).

[114] Wyrick, J. J., Aparicio, J. G., Chen, T., Barnett, J. D., Jennings, E. G., Young, R. A., Bell, S. P. & Aparicio, O. M. Genome-wide distribution of ORC and MCM proteins in *S. cerevisiae*: high-resolution mapping of replication origins. *Science* **294**:2357–23650 (2001).

[115] Knott, S. R. V., Viggiani, C. J., Tavaré, S. & Aparicio, O. M. Genome-wide replication profiles indicate an expansive role for Rpd3L in regulating replication initiation timing or efficiency, and reveal genomic loci of Rpd3 function in *Saccharomyces cerevisiae*. *Genes Dev.* **23**:1077–1090 (2009).

[116] Wu, P.-Y. J. & Nurse, P. Establishing the program of origin firing during S phase in fission yeast. *Cell* **136**:852–864 (2009).

[117] Campbell, N. A. & Reece, J. B. *Biology* (Pearson Benjamin Cummings, San Francisco, CA, 2008), 8th ed.

[118] Henikoff, S. Nucleosome destabilization in the epigenetic regulation of gene expression. *Nat. Rev. Genet.* **9**:15–26 (2008).

[119] Eberharter, A. & Becker, P. B. Histone acetylation: a switch between repressive and permissive chromatin. *EMBO Rep.* **3**:224–229 (2002).

[120] Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**:289–293 (2009).

[121] Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., Bell, G. W., Walker, K., Rolfe, P. A., Herbolsheimer, E., Zeitlinger, J., Lewitter, F., Gifford, D. K. & Young, R. A. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**:517–527 (2005).

[122] Arneodo, A., Vaillant, C., Audit, B., Argoul, F., d‘Aubenton Carafa, Y. & Thermes, C. Multi-scale coding of genomic information: From DNA sequence to genome structure and function. *Phys. Rep.* **498**:45–188 (2011).

[123] Goldar, A., Labit, H., Marheineke, K. & Hyrien, O. A dynamic stochastic model for DNA replication initiation in early embryos. *PLoS ONE* **3**:e2919 (2008).

[124] Chen, G.-H., Tang, J. & Leng, S. Prior image constrained compressed sensing (PICCS): A method to accurately reconstruct dynamic CT images from highly undersampled projection data sets. *Med. Phys.* **35**:660–663 (2008).

[125] Fried, J. Method for the quantitative evaluation of data from flow microfluorometry. *Compt. Biomed. Res.* **9**:263–276 (1976).

[126] Gray, J. W. Cell-cycle analysis of perturbed cell populations: computer simulations of sequential DNA distributions. *Cell Proliferation* **9**:499–516 (1976).

[127] Höcker, A. & Kartvelishvili, V. SVD approach to data unfolding. *Nucl. Inst. Meth. A* **372**:469–481 (1996).

[128] Grant, M. & Boyd, S. CVX: Matlab software for disciplined convex programming, version 1.21. `../../cvx` (2011).

[129] Grant, M. & Boyd, S. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*, eds. Blondel, V., Boyd, S. & Kimura, H., Lecture Notes in Control and Information Sciences, pp. 95–110 (Springer-Verlag Limited, 2008). `http://stanford.edu/~boyd/graph_dcp.html`.

[130] Boyd, S. & Vandenberghe, L. *Convex Optimization* (Cambridge University Press, New York, NY, 2004).

[131] Byrd, R. H., Gilbert, J. C. & Nocedal, J. A trust region method based on interior point techniques for nonlinear programming. *Math. Programming* **89**:149–185 (2000).

[132] Wächter, A. & Biegler, L. T. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Programming* **106**:25–57 (2006).

[133] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B* **58**:267–288 (1996).

[134] Candès, E. J. & Wakin, M. B. An introduction to compressive sampling. *IEEE Signal Processing Mag.* **25**:21–30 (2008).

[135] Marheineke, K. & Hyrien, O. Aphidicolin triggers a block to replication origin firing in *Xenopus* egg extracts. *J. Biol. Chem.* **276**:17092–17100 (2001).

[136] Bechhoefer, J. & Marshall, B. How *Xenopus laevis* replicates DNA reliably even though its origins of replication are located and initiated stochastically. *Phys. Rev. Lett.* **98**:098105 (2007).

[137] Kimelman, D., Kirschner, M. & Scherson, T. The events of the midblastula transition in *Xenopus* are regulated by changes in the cell cycle. *Cell* **48**:399 (1987).

[138] Blow, J. J., Gillespie, P. J., Francis, D. & Jackson, D. A. Replication origins in *Xenopus* egg extract are 5–15 kilobases apart and are activated in clusters that fire at different times. *J. Cell Biol.* **152**:15–26 (2001).

[139] O. Hyrien (private communication).

[140] Prokhorova, T. A., Mowrer, K., Gilbert, C. H. & Walter, J. C. DNA replication of mitotic chromatin in *Xenopus* egg extracts. *Proc. Natl. Acad. Sci.* **100**:13241–13246 (2003).

[141] Hensey, C. & Gautier, J. Programmed cell death during *Xenopus* development: A spatio-temporal analysis. *Dev. Biol.* **203**:36–48 (1998).

[142] Walter, J. & Newport, J. W. Regulation of replicon size in *Xenopus* egg extracts. *Science* **275**:993 – 995 (1997).

[143] Harvey, K. J. & Newport, J. CpG methylation of DNA restricts prereplication complex assembly in *Xenopus* egg extracts. *Mol. Cell Biol.* **23**:6769–6779 (2003).

[144] Laskey, R. A. Chromosome replication in early development of *Xenopus laevis*. *J. Embryol. Expe. Morphol. Suppl.* **89**:285–294 (1985).

[145] Romanowski, P., Madine, M. A., Rowles, A., Blow, J. & Laskey, R. A. The *Xenopus* origin recognition complex is essential for DNA replication and MCM binding to chromatin. *Curr. Biol.* **6**:1416–1425 (1996).

[146] Rhind, N. DNA replication timing: random thoughts about origin firing. *Nat. Cell Biol.* **8**:1313–1316 (2006).

[147] Lucas, I., Chevrier-Miller, M., Sogo, J. M. & Hyrien, O. Mechanisms ensuring rapid and complete DNA replication despite random initiation in *Xenopus* early embryos. *J. Mol. Biol.* **296**:769–786 (2000).

[148] 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.* **279**:28071–28081 (2004).

[149] 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* **18**:3059–3067 (2007).

[150] Thiébaud, C. H. & Fischberg, M. DNA content in the genus *Xenopus*. *Chromosoma* **59**:253–257 (1977).

[151] Bell, S. P. & Dutta, A. DNA replication in eukaryotic cells. *Annu. Rev. Biochem.* **71**:333–374 (2002).

[152] Kimmel, M. & Axelrod, D. E. *Branching Processes in Biology* (Springer-Verlag, New York, NY, New York, 2002).

[153] Spall, J. C. *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control* (J. Wiley, Hoboken, NJ, New Jersey, 2003).

[154] Pontarin, G., Fijolek, A., Pizzo, P., Ferraro, P., Rampazzo, C., Pozzan, T., Thelander, L., Reichard, P. A. & Bianchi, V. Ribonucleotide reduction is a cytosolic process in mammalian cells independently of DNA damage. *Proc. Natl. Acad. Sci.* **105**:17801–17806 (2008).

[155] Postow, L., Crisona, N. J., Peter, B. J., Hardy, C. D. & Cozzarelli, N. R. Topological challenges to DNA replication: Conformations at the fork. *Proc. Natl. Acad. Sci.* **98**:8219–8226 (2001).

[156] Skogestad, S. & Postlethwaite, I. *Multivariable Feedback Control* (J. Wiley, New York, NY, Chichester, 2005), 2nd ed.

[157] Karschau, J., Blow, J. J. & de Moura, A. P. S. Optimal placement of origins for DNA replication. *Phys. Rev. Lett.* **108**:058101 (2012).

[158] Meister, P., Taddei, A., Ponti, A., Baldacci, G. & Gasser, S. M. Replication foci dynamics: replication patterns are modulated by S-phase checkpoint kinases in fission yeast. *EMBO J.* **26**:1315–1326 (2007).

[159] Canman, C. E. Replication checkpoint: preventing mitotic catastrophe. *Curr. Biol.* **11**:R121–R124 (2001).

[160] Piccoli, G. D., Katou, Y., Itoh, T., Nakato, R., Shirahige, K. & Labib, K. Replisome stability at defective DNA replication forks is independent of S phase checkpoint kinases. *Mol. Cell* **45**:696–704 (2012).

[161] Chen, K. C., Calzone, L., Csikasz-Nagy, A., Cross, F. R., Novak, B. & Tyson, J. J. Integrative analysis of cell cycle control in budding yeast. *Mol. Biol. Cell* **15**:3841–3862 (2004).

[162] Rhind, N. & Gilbert, D. M. *DNA Replication*, chap. Replication Timing (Cold Spring Harbor Press, 2012, in press), 3rd ed.

[163] Farkash-Amar, S., Lipson, D., Polten, A., Goren, A., Helmstetter, C., Yakhini, Z. & Simon, I. Global organization of replication time zones of the mouse genome. *Genome Res.* **18**:1562–1570 (2008).

[164] Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T. C., Robins, A. J., Dalton, S. & Gilbert, D. M. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res.* **20**:761–770 (2010).

[165] Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., Blau, C. A. & Noble, W. S. A three-dimensional model of the yeast genome. *Nature* **463**:363–367 (2010).

[166] Lu, J., Li, F., Murphy, C. S., Davidson, M. W. & Gilbert, D. M. G2 phase chromatin lacks determinants of replication timing. *J. Cell Biol.* **189**:967–980 (2010).

[167] Cohen, J. E. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biol.* **2**:e439 (2004).

[168] Phillips, R. A vision for quantitative biology. i-Bio-Magazine (2010). `http://ibiomagazine.org/index.php/issues/august-issue/quantitative-biology`.