Modelling an optical replication mapping experiment to measure human DNA replication kinetics

by

Sina Falakian

B.Sc. (Physics), Sharif University of Technology, 2020

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
Department of Physics
Faculty of Science

SIMON FRASER UNIVERSITY
Spring 2024

Declaration of Committee

Name:	**Sina Falakian**
Degree:	**Master of Science**
Thesis title:	**Modelling an optical replication mapping experiment to measure human DNA replication kinetics**
Committee:	**Chair:** Nancy Forde Professor, Physics
	John Bechhoefer Supervisor Professor, Physics
	Eldon Emberly Committee Member Professor, Physics
	Andrei Frolov Examiner Professor, Physics

Abstract

DNA replication is nature's method of copying genetic information, a crucial process facilitated by a set of protein machinery at specific sites called "replication origins." The spatiotemporal organization of these origins, known as the "replication program," governs DNA replication timing, a topic of longstanding interest in biology. Various experimental approaches, such as DNA combing and DNA sequencing, have been employed to measure the replication program, particularly in bacterial and simple eukaryotic systems. However, these methods face challenges when applied to human genomes, owing to their longer length and higher stochasticity.

Optical Replication Mapping (ORM) has emerged as a novel experimental approach capable of providing single-molecule, genome-wide, and high-throughput data on the replication process. Nonetheless, ORM experiments suffer from sparse labeling, necessitating a model to understand label distribution around origins. In this thesis, we show that a previously used model for label incorporation is incomplete. In response, we have refined the model by introducing further physical assumptions. Despite these refinements, the more elaborate models still fall short of explaining ORM data comprehensively.

Keywords: DNA replication; Optical Replication Mapping; Initiation event; Label incorporation

Table of Contents

List of Tables

List of Figures

Deoxyribonucleic acid, or DNA, functions as a polymer containing genetic information crucial for the development, functioning, and reproduction of all known living organisms and many viruses [1, 2]. DNA replication, the process of copying DNA molecules prior to cell division, is vital for transmitting genetic information across generations. In eukaryotic organisms, including humans, DNA replication begins by "initiation events" or firing "replication origins," where two DNA strands are separated. The spatiotemporal organization of the initiation events is called the "replication program" and determines the replication timing of different cells.

In the early 2000s, the development of new experimental techniques made genome-wide studies of DNA replication possible [3]. These experimental methods inspired new techniques to quantify DNA replication by translating experimental data into quantitative models. Bacterial chromosomes have a single, sequence-specific origin, whereas eukaryotic organisms have multiple replication origins [1]. Additionally, DNA replication in archaea species lies in between these two cases because some species have a single chromosome with a single origin, whereas others have multiple origins per chromosome [4]. These studies also show that eukaryotic replication is stochastic [5, 6]. For example, in *S. cerevisiae*, origins are at specific sites along the genome, but these sites fire randomly during each replication realization [7]. In *Xenopus laevis* embryos, the initiation positions are distributed with a homogeneous probability density over the genome [8]. In somatic metazoan cells, there is stochastic clustering of initiation events over the genome [9]. Here, the name "initiation event" becomes more useful, since there is no specific site for initiations, whereas "origin" suggests specific sites [10].

Unraveling the complexity of replication program in human genomes remains a significant challenge [9]. The challenge arises from the inefficiency and heterogeneity observed in the firing of replication origins in mammalian cells [6]. A groundbreaking experimental method, Optical Replication Mapping (ORM), was introduced in 2019 [10] to overcome the shortcomings of previous experimental approaches. In this thesis, we analyze ORM data to gain deeper insights into the experiment and infer the kinetics underlying DNA replication.

In Chapter 1, we provide an overview of the biology, modeling approaches, and prior experiments related to DNA replication. Subsequently, in Chapter 2, we present the numerical methods employed to simulate experimental data and test our models. Finally, in Chapter 3, we introduce new models and compare their predictions against the experimental results.

Chapter 1

Introduction

In this chapter, we offer an overview of DNA structure and the intricacies of DNA replication in Section 1.1. In Section 1.2, we review the experimental approaches used to investigate DNA replication kinetics and introduce ORM as a new approach to provide higher-throughput data for genome-wide replication kinetics. Finally, the mathematical framework employed in this research for modeling DNA replication kinetics is presented in Section 1.3.

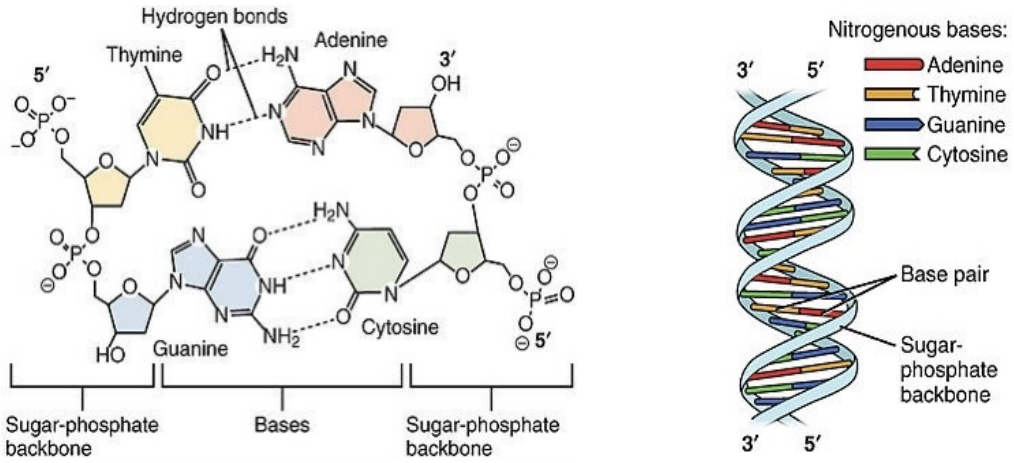
1.1 Biology background

1.1.1 DNA structure

Under typical physiological conditions, deoxyribonucleic acid (DNA) has a double-helix structure. Its two strands are each composed of four types of nucleotide: deoxyribose sugar, phosphate group, and one of four nitrogenous bases (adenine, thymine, cytosine, or guanine) [1, 11] (see Fig. 1.1). These strands are held together by hydrogen bonds between nucleotide pairs, where adenine pairs with thymine and cytosine pairs with guanine, forming complementary base pairs. The sequence of these base pairs forms the genetic code, which cellular machinery reads and translates into proteins.

Crucially, DNA strands exhibit specific directionality due to the arrangement of carbon atoms in the deoxyribose sugar. One strand runs from $5'$ to $3'$, while the other runs from $3'$ to $5'$ (see Fig. 1.1). This directional orientation plays a vital role in DNA replication and transcription processes. Within eukaryotic cells, DNA mostly resides within the nucleus, organized into chromosomes composed of DNA and histone proteins[1]. Organisms with diploid genomes possess homologous chromosomes containing similar genes but potentially different alleles. In humans, there are typically 23 pairs of chromosomes.

[1]DNA also exists in mitochondria [12].



(a) (b)

Figure 1.1: **DNA structure.** a): Molecular structure of DNA, comprising four bases and the sugar-phosphate backbone. Paired bases form hydrogen bonds. (b) Double-helical structure of DNA, consisting of two DNA strands with opposite directions (3′-5′ and 5′-3′) [13].

1.1.2 DNA replication

DNA replication is the fundamental process by which a cell duplicates its DNA (parental DNA) to generate identical copies (daughter DNA) for distribution to its daughter cells during cell division [14, 15]. This critical event occurs during the S phase (synthesis phase) of the cell cycle, which is an integral part of this process. The cell cycle encompasses a series of events leading to cell division and duplication, comprising distinct phases, each with specific functions and characteristics.

In somatic human cells, the cell cycle comprises four phases [1], as illustrated in Fig. 1.2:

- **G1 Phase (Gap 1):** The cell undergoes growth and prepares for DNA replication.
- **S Phase (Synthesis):** DNA replication takes place, ensuring the genetic material is duplicated to provide each daughter cell with an identical set of chromosomes.
- **G2 Phase (Gap 2):** Following DNA replication, the cell continues to grow and readies itself for division. Additional proteins and organelles are synthesized to support the upcoming cell division.
- **Mitotic Phase (M phase):** The cell undergoes division, resulting in the formation of two daughter cells.

In the process of DNA replication, the original DNA is referred to as “parental DNA,” and the resulting two replicated DNAs are known as “daughter DNAs” [1, 2]. The replication process initiates with the unwinding of the two DNA strands. Subsequently, DNA synthesis occurs on these separated DNA strands, leading to the formation of two complementary

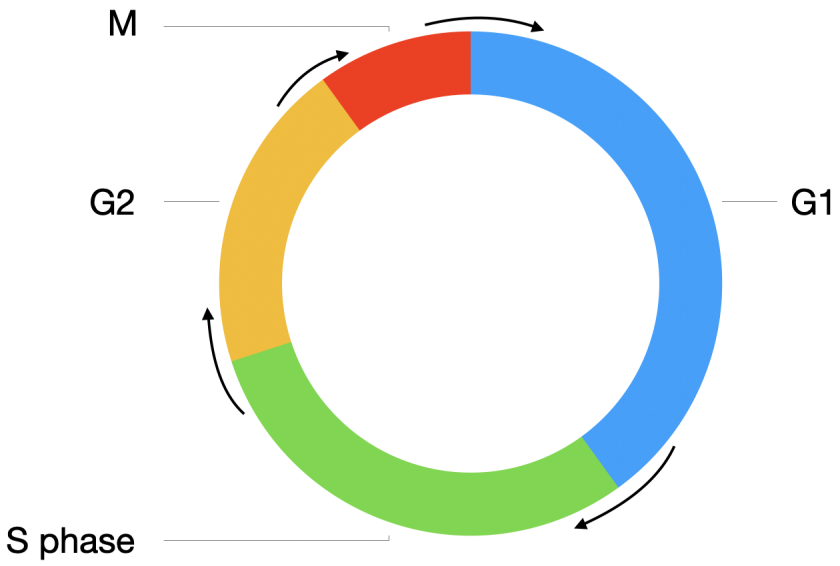


Figure 1.2: **Somatic eukaryotic cell cycle.** G1: Growth, S: DNA synthesis, G2: Growth and preparation for mitosis, M: Mitosis (cell division). The duration of these cell cycle phases varies considerably in different kinds of cells. For a typical rapidly proliferating human cell with a total cycle time of 24 hours, the G1 phase might last about 11 hours, S phase about 8 hours, G2 about 4 hours, and M about 1 hour.

DNA strands, each pairing with a parental DNA strand. As a result, two daughter DNAs are produced, with each sharing one strand with the parental DNA.

This unwinding of the two strands is termed an "initiation event," and the specific position along the genome where initiation occurs is referred to as an "initiation position." These initiation positions are commonly known as "replication origins," and the act of initiation is often described as the "firing" of these origins.

The DNA replication process consists of four stages [15]: licensing, initiation, elongation, and termination.

- **Licensing:** In the G1 phase, several proteins bind to various locations along the genome, creating potential initiation sites [16, 15, 17]. During this process, proteins comprising the pre-replication complex (pre-RC) assemble at these potential replication initiation locations along the genome, as depicted in Fig.1.3. The pre-RC includes essential proteins such as the origin recognition complex (ORC), Cdc6, and Cdt1. Proteins Cdc6 and Cdt1 play pivotal roles by recruiting and loading the MCM complex onto the DNA at replication origins after licensing is completed. Subsequently, the MCM complex unwinds and separates the DNA strands, facilitating fork elongation.

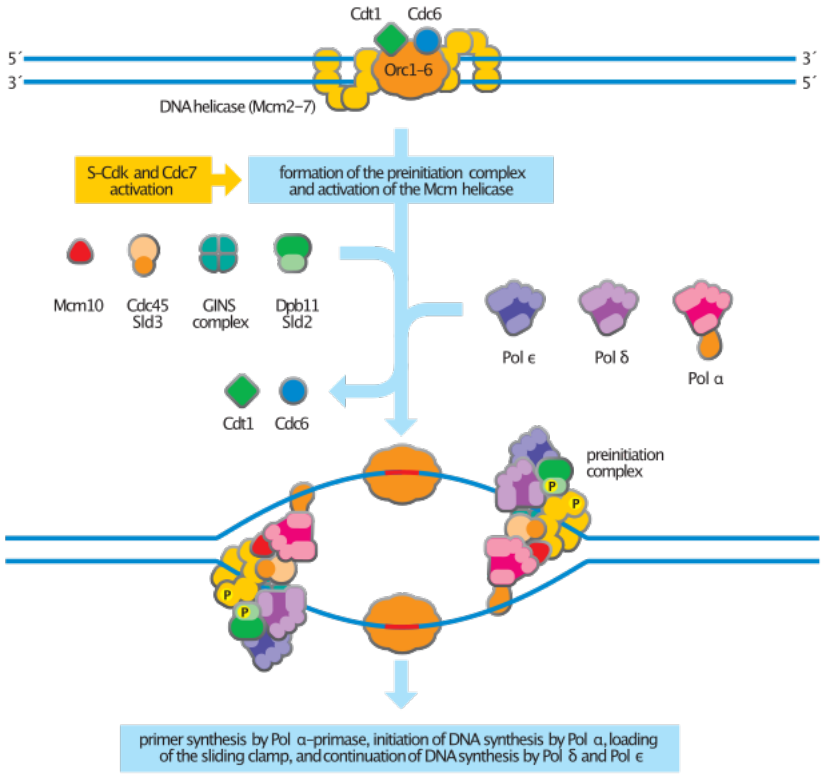


Figure 1.3: **Pre-replication complexes bind to DNA as potential initiations.** Pre-RC consists of ORC and two MCM proteins. This complex is activated by recruiting additional proteins that unbind the two DNA strands. The two MCM rings along with some additional proteins form helicase [18].

- **Initiation:** DNA replication commences with initiation events, where a subset of potential origins is activated and the two DNA strands unwind at specific initiation positions along the genome [19]. After the licensing process is completed and S phase begins, certain pre-replication complexes (pre-RCs) are activated through the phosphorylation by cyclin-dependent kinase (CDK) and Dbf4-dependent kinase (DDK), along with other protein factors [1, 17]. These proteins collectively form a pre-initiation complex (Pre-IC). Through a sequence of molecular reactions, the two DNA strands are unwound at the Pre-IC, marking an initiation event or origin firing, as illustrated in Fig. 1.3.

  During an initiation event, the combination of Cdc45, MCM2-7, and GINS (CMG complex) forms a "helicase" enzyme, responsible for unwinding and separating the two DNA strands. The region where the two DNA strands are separated is referred to as a "replication bubble," and the point where the replication bubble intersects with the double-stranded DNA (dsDNA) is termed a "replication fork" [20, 17] (see Fig. 1.4). Each initiation event gives rise to two replication forks that move away

from the initiation position. As these forks progress, the helicase enzyme unzips the dsDNA, causing the replication bubble to expand.

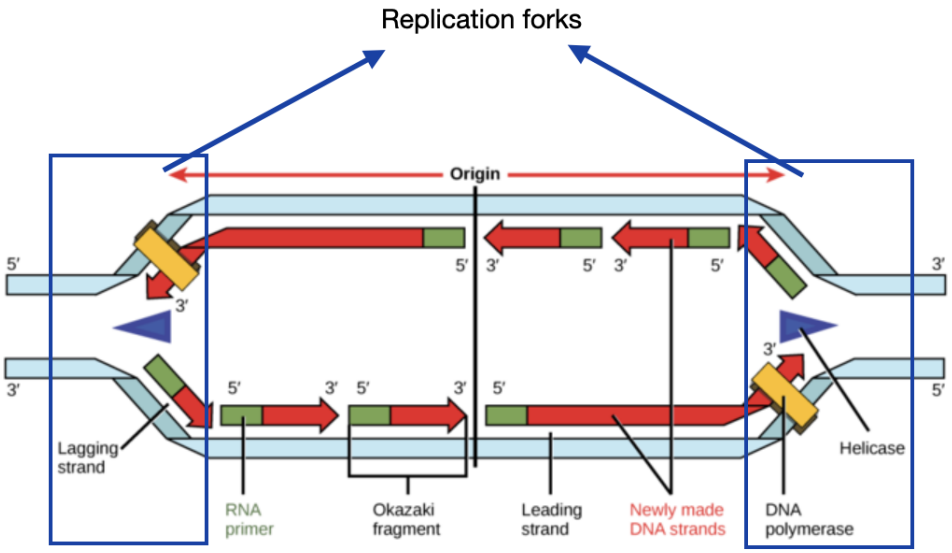


Figure 1.4: **DNA Replication bubble.** Complementary DNA strands consist of leading and lagging strands. The leading strand is a single fragment growing as the replication forks move. Lagging strands consist of multiple Okazaki fragments. All fragments (red) are initiated by RNA primers (green). Image downloaded and modified from [21].

- **Elongation:** DNA replication elongation is the process by which new DNA strands are synthesized using the existing DNA strands as templates [20, 17]. This synthesis occurs in the $5'$ to $3'$ direction, signifying that DNA is built by adding nucleotides to the growing strand's $3'$ end. To initiate DNA synthesis, a short RNA primer is synthesized by an enzyme called primase. This RNA primer serves as the starting point for DNA polymerases, enabling them to bind free nucleotides within the nucleus to the single-stranded DNA (ssDNA), as depicted in Fig. 1.4.

  During replication elongation, the two separated DNA strands are synthesized in distinct directions [1]. The leading strand, which is continuously synthesized in the same direction as the movement of the replication fork, is replicated by DNA polymerase epsilon. This polymerase continuously adds nucleotides to the growing $3'$ end of the leading strand, forming a complementary DNA strand [4]. Conversely, the lagging strand is synthesized discontinuously in the opposite direction of the replication fork movement, resulting in short "Okazaki" fragments [22]. DNA polymerase delta is responsible for synthesizing these fragments. As the replication fork continues to open, primase generates new RNA primers on the lagging strand, and DNA polymerase delta synthesizes short DNA stretches from each primer.

- **Termination:** DNA replication termination signifies the ultimate stage in the DNA replication process, where the replication forks converge, and the synthesis of the new DNA strands concludes [11, 23] (see Fig. 1.5). After the termination process is finished, DNA is copied, and the cell enters G2 phase.

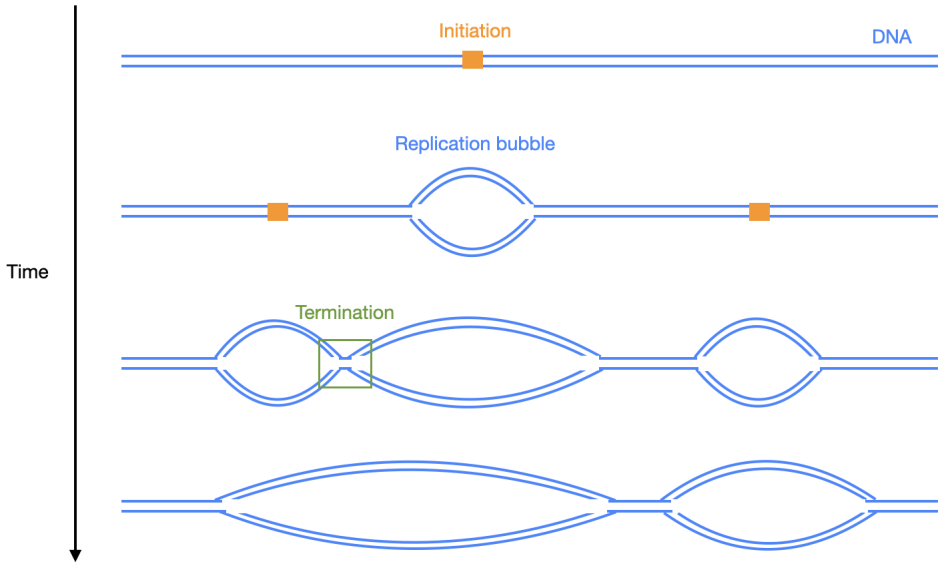


Figure 1.5: **DNA Replication in eukaryotes.** Multiple initiation events occur at different times during S phase and replication bubbles merge into larger bubbles at termination sites.

The description of the replication process outlined above can exhibit variations in detail among different cell types; however, these four stages are typically observed in somatic cells [1]. Key determinants of replication kinetics include initiation positions, times, and fork velocity. These factors collectively constitute the "replication program" [7]. In Section 1.3, we will describe quantitative models for replication kinetics.

The replication program plays a pivotal role in determining replication timing [5]. DNA replication timing refers to the temporal order and coordination of DNA replication at various genomic regions throughout the cell cycle. It determines when and where (stochastically) specific segments of DNA molecules are replicated within a cell's nucleus. Notably, the overall replication timing can vary significantly across different cell types, as illustrated in Table 1.1.

In E. coli, DNA replication initiates from a specific site known as oriC [24]. However, in metazoan cells, the replication program is significantly more complex [5] (see Fig. 1.5). Eukaryotes have longer genome size and slower fork velocity than do bacteria (see Table 1.1). As a result, they have multiple initiations to regulate the total replication timing. These multiple initiation events occur at various points across the genome to regulate the overall replication timing.

Previous studies examining biological data on DNA replication in *S. cerevisiae* have identified sequence-specific initiation sites (about 700) on the yeast genome [25]. In each replication realization, approximately one-third of these initiation sites are activated [26]. The rate at which these origins fire is stochastic both in position and time. The efficiency of each initiation site, denoted as the fraction of replication realizations where an initiation site fires, can be estimated based on the number of MCMs loaded at the initiation site [27].

In contrast to simpler metazoan organisms, precisely pinpointing the locations and temporal activation patterns of DNA replication origins in humans remains a challenge [9]. This challenge primarily arises from the inefficiency and heterogeneity observed in the firing of replication origins in mammalian cells [6]. Moreover, in more common realizations, origins appear to initiate within broader initiation zones (IZs) spanning tens of kilobases [28]. Within these zones, the firing efficiency is notably lower, often less than 1%. This complexity presents a significant barrier to fully understanding the dynamics and regulation of DNA replication in human cells [29].

	Escherichia Coli	*Saccharomyces Cerevisiae*	*Homo Sapiens*
DNA length (bp)	4.6×10^6	24×10^6	6.4×10^9
Fork Velocity (bp/s)	1000	30	30
Total replication time (min)	25	40	500
Average number of initiations	1	500	50000

Table 1.1: **Replication program in different cell types [11].**

In this section, we have presented information about the DNA replication program, covering molecular mechanisms and replication programs in higher eukaryotes. Below, we review the experimental methods used to measure the replication program (see Section 1.2) and explore modeling approaches employed to quantify replication kinetics based on experimental data in Section 1.3.

1.2 Experimental techniques for studying DNA replication

A wide array of experiments has been conducted to measure replication program and kinetics in human genomes [3]. While earlier methods provided valuable insights into DNA replication, they fell short of mapping genome-wide replication kinetics comprehensively (see Subsection 1.2.1). Since the late 90s, new methods emerged to map genome-wide replication kinetics, experiments on cell population level to single-molecule and single-cell. These experiments were conducted on synchronized cells, on various fractions of cells sorted through the S phase, and on asynchronous cell cultures. While most techniques generated data post-replication, some methods allowed for the measurement of live replication dynamics.

In this section, we introduce two broad categories of experiments based on their data: replication timing, which estimates the replication time for each position along the genome (see Subsection 1.2.2), and single-fiber methods, which analyze data derived from labeled

DNA fibers (see Subsection 1.2.3). These experiments led to the inference of the genome-wide replication program in simple eukaryotes and the identification of regions with a higher probability of initiation in metazoan genomes (see Section 1.3). However, an accurate measurement of the replication program in metazoan genomes remains elusive. To address the limitations of previous experimental techniques, Optical Replication Mapping (ORM) was developed [10]. We will discuss this experiment in Subsection 1.2.4.

1.2.1 Early experiments

The first attempt to understand DNA replication is the Meselson-Stahl experiment [30] from 1958. The authors used nitrogen isotopes and found results that support a semi-conservative model of DNA replication, where each strand of the double helix serves as a template for the synthesis of a new complementary strand.

One of the earliest single-molecule approaches, dating back to 1963, used autoradiography, a technique that employed radioactive labeling to track DNA replication [31]. Developed by John Cairns, this method involved using the radioactive isotope ^{3}H-thymine, which replaces regular thymine during replication. The labeled DNA is then extracted and denatured into single strands, followed by spreading DNA fibers on an emulsion, a technique that separates DNA molecules based on size.

Starting in the 1970s, researchers used two-dimensional gel electrophoresis (2D gel electrophoresis) to explore DNA replication dynamics [32]. This technique involved sorting DNA fragments based on their charge and size in two dimensions within gels. The separated DNA fragments were then visualized by Southern hybridization. These experiments provided local information about the replication kinetics but did not offer genome-wide information about replication kinetics.

After electrophoresis, the gel is placed in close contact with X-ray film or a phosphor imaging plate [32]. The radioactive isotopes in the DNA emit radiation, and the regions on the film exposed to this radiation darken, forming an image that represents the distribution and intensity of DNA synthesis. Autoradiography provides valuable information about DNA replication rate, the presence of replication forks, and the fidelity of DNA synthesis. However, it does not provide enough information to infer the replication program accurately.

These early experiments failed to provide genome-wide information about replication kinetics. Starting in the 1990s, several innovative experimental approaches were developed to provide genome-wide data [3]. These datasets can be classified in two categories: DNA replication profiles, which provide average timing data of replication kinetics over the genome, and single-fibre approaches, where the data is collected from individual DNA fibres. Each of these two categories has their advantages and disadvantages, which we will discuss in the next two subsections.

1.2.2 DNA replication timing profile

DNA replication timing refers to the temporal order in which different regions of a genome replicate during the cell cycle [3]. It is a highly regulated process, and the timing of DNA replication can vary among different cell types and developmental stages. The replication timing profile of a genome can have important implications for cellular function, stability, and gene expression (Fig. 1.6). The height of DNA replication timing profiles indicate the average time when each position along the genome is replicated. Several techniques have been developed to study DNA replication timing profiles, and advances in genomic technologies have allowed for more detailed analyses. The height of the profile represented the replication timing across the chromosome, with peaks indicating replication origin sites.

One of the primary assays for genome-wide replication timing in mammalian cells has been "Repli-Seq," in which cells labeled with the thymine analogue Bromodeoxyuridine (BrdU) during 10–20% of S phase are sorted into early and late S fractions, and replication timing profiles are generated from the log ratio of read enrichment in the BrdU-immunoprecipitated early fraction to that in the late fraction [33]. Enrichment numbers are generated either by microarray hybridization [34] or, more recently, by DNA sequencing [35]. This approach has also been applied to single cells to infer cell-to-cell variation in replication timing profile [36]. But this method suffers from low resolution because of the limited coverage of whole genome single-cell sequencing and the single-time snapshot obtained from each cell.

Another method involves sorting cells through S phase, where cells are labeled with fluorescent dyes and then categorized into multiple fractions based on their fluorescent content [37]. After sorting, copies are made of DNA segments at different time points through the S phase [3, 38]. A replicated segment of DNA generates two times the amount of DNA as an unreplicated segment; thus, the copy number of a segment of DNA in a cell culture can vary between the total number parental DNA and twice this number. This allows for the calculation of the replication timing profile based on variations in copy number.

The copy-number approach was extended to single-cell analyses, offering a more detailed understanding of DNA replication dynamics at the single-cell level [39, 40, 41]. These advances provided valuable insights into replication timing and contributed to the development of the "replication domain" model, where extended segments spanning 400 kb to 1.2 Mb exhibited uniform replication timing, leading to relatively even plateaus on replication profiles [42]. Studies show significant correlations between replication domains and the spatial structure of DNA inside the nucleus, coming from substructure within chromatin compartments called topologically associating domains (TADs) [43]. TADs are regions of the genome where DNA sequences are more likely to interact with each other spatially than with sequences outside the domain. Also, studying replication timing in different species shows that the temporal program evolves with species, but the variations along the genome are relatively fixed [44]. Association of replication timing variation with genetic

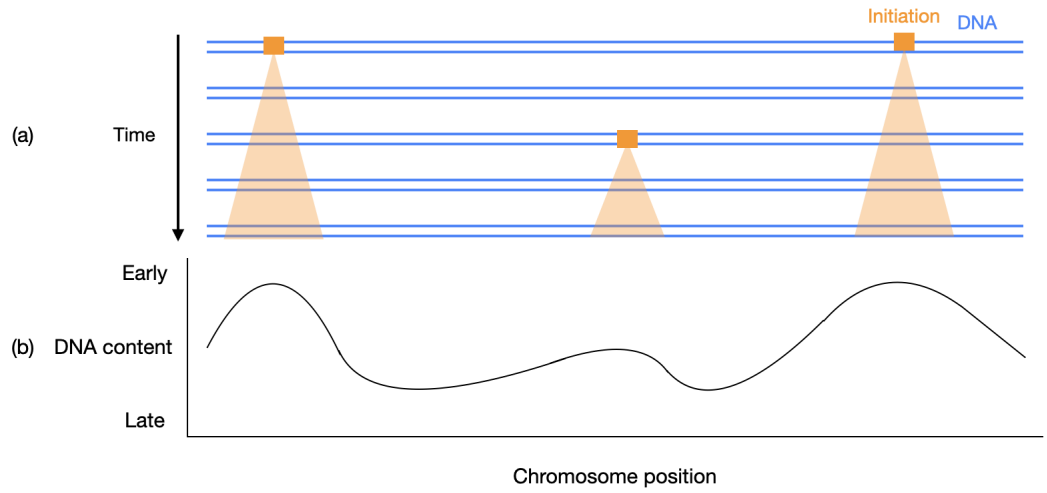


Figure 1.6: **Replication timing profile.** (a) The replicated regions of DNA (orange) grow during S phase. The regions closer to initiation positions replicate earlier. (b) The DNA content is a measure for the replication timing of each segment of a chromosome. Higher peaks demonstrate regions with higher probability of initiation.

variation revealed that DNA sequence evolution can explain replication timing variation between species. These interactions are thought to be facilitated by the looping of DNA [45]. Although replication timing has provided valuable information about the genome-wide replication program, inferences are at the scale above the replication domain sizes (minimum 400 kb). For inferring the replication program with higher resolution, single-fibre experiments were developed, which will be discussed in the next subsection.

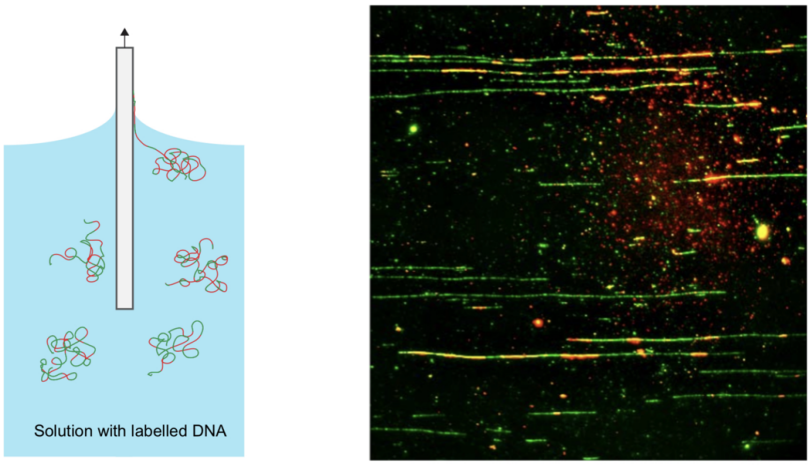
1.2.3 Single-fibre experiments

Single-fibre experiments include approaches where the replication program is inferred by analyzing the data coming from individual DNA fibres. These methods provide more accurate information about replication kinetics at different sites along the genome than methods based on cell-population data. In single-fibre experiments, DNA molecules are labeled with nucleotide analogues at different time points during the replication process [3]. Subsequently, the replicated DNA fragments are visualized and replication tracks become visible due to the incorporation of these labeled nucleotides.

A traditional approach is fibre autoradiography (as mentioned in Subsection 1.2.1). Since the 1990s, more powerful solutions to the problems of heterogeneity and low SNRs were developed based on single-molecule analysis, which allows the identification of sites of replication initiation on individual DNA fibers [46]. Other single-molecule approaches include DNA analysis with fluorescence in situ hybridization (FISH) using sequence-specific probes, such as the SMARD (single molecule analysis of replicating DNA) technique [47]. These techniques have provided critical insight into the location and firing kinetics of

mammalian replication origins. However, they are restricted to the analysis of at most a few genomic loci. Nanopore sequencing-based methods have been employed in the study of DNA replication [48], offering unique advantages in capturing real-time dynamics. However, current experiments do not allow genome-wide analysis and long range reads more than 30 kb. Another real-time approach uses rolling-circle DNA amplification scheme to map leading strand synthesis in stretched *E. coli* DNA [49].

Starting in 1997, DNA combing was developed, which is a technique for stretching DNA fibres on combing glass. It has been used for high-resolution studies of genomes such as human [50]. Labeling the DNA fibres during the replication process with nucleotide analogues allowed researchers to study genome-wide replication programs both in yeast and mammalian cells [51, 52]. This approach provided high-resolution, single-molecule data of DNA fibres, which provided new insight about replication program. In DNA combing, researchers label newly synthesized DNA strands with nucleotide analogs or fluorescent markers such as BrdU or Ethynyldeoxyuridine (EdU) (see Fig. 1.7) [50, 3]. These labels are incorporated into the replication tracks. DNA fibres then stick onto the surface of combing glass, where they are stretched out linearly. The surface is treated to ensure that the ends of DNA fibre adhere preferentially, which aids linear stretching (Fig. 1.7). Subsequently, the DNA molecules are denatured on the glass, and fluorescent dyes are attached to the labels. Using fluorescence microscopy, researchers can visualize the labeled DNA strands along their length.



(a) (b)

Figure 1.7: **DNA combing experiment.** (a) DNA fibres in solution bind to the surface of the combing glass and stretched by the meniscus force of water. (b) Fluorescent image taken from a combing glass after taking out the combing glass. The green shows the DNA fibres and red are the nucleotide analogues (green and red mix to show yellow). Courtesy of Dr. Nicholas Rhind.

Identifying the positions of labeled sites on the DNA strands, gives valuable information about the movement of replication forks, which are the points where DNA replication is actively taking place. DNA fibres can be localized along the genome using fluorescent probes binding to specific sequences called restriction sites along the genome (barcoding) [53]. Analyzing the distances between these labeled sites allows one to deduce the speed of replication forks and determine the timing of replication initiation and termination. However, another challenge with this set of experiments is low-throughput data. Because a significant portion of the DNA fibres that adhere to the glass surface is not adequately stretched, one cannot accurately analyze these strands after visualization. Consequently, the data obtained from these experiments is low throughput. This limits the ability to infer genome-wide information about replication program in human genomes.

To overcome the shortcomings of previous experiments, Optical Replication Mapping was developed as a new single-molecule high-throughput approach to map genome-wide replication kinetics in human genomes [54, 10]. Optical Replication Mapping technology enables analysis of long DNA fibres (150 kb to 2 Mb) [55]. Using nucleotide-analog incorporation, this approach can be used to study DNA replication kinetics [54]. However, previous approaches rely on DNA extracted from cells. In the next section, we present a new ORM experiment done in vivo and discuss how to understand the data coming from this experiment [10].

1.2.4 Optical Replication Mapping

Optical Replication Mapping (ORM) is a single-molecule, high-throughput, in vivo approach, providing post-replication data of replication kinetics in the human genome. This experiment is conducted in both synchronous and asynchronous cell cultures.

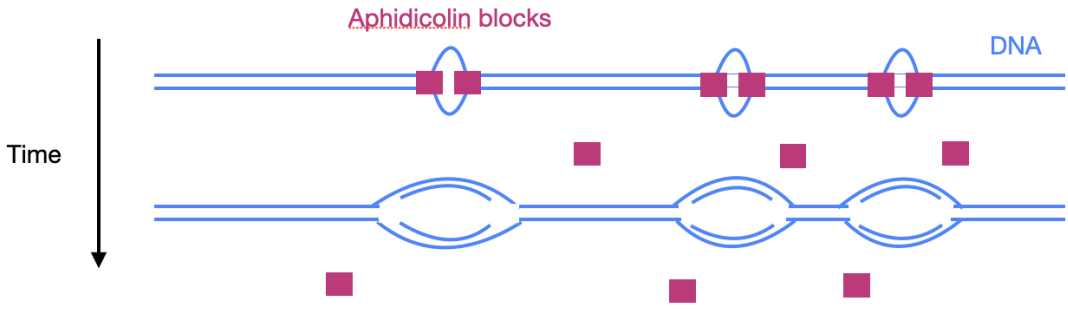


Figure 1.8: **Arresting replication forks using aphidicolin blocks.** Aphidicolin blocks freeze replication forks right after initiation. Elongation starts right after aphidicolin is released.

In this experiment, HeLa S3 cells, a widely used human cancer cell line, are synchronized using aphidicolin blocks (Fig. 1.8). Aphidicolin blocks are natural inhibitors of DNA synthesis that target DNA polymerases, especially DNA polymerase alpha (Pol α) and DNA polymerase

delta (Pol δ) [56]. They are added to a culture of cells excluding the cells in their S phase (see Fig. 1.2). Thus, the fork progression is stopped for all the cells at the beginning of S phase and all the cells will be synchronized at the begging of the S phase. After a period of cell synchronization with aphidicolin, the block can be released by removing the inhibitor. This allows DNA replication to resume. This process impacts the replication dynamics by increasing the number of initiations occurring at time $t = 0$. We will discuss this further in Chapter 3.

After synchronization, the cell is electroporated with fluorescent deoxyuridine-triphosphate (dUTP) molecules. In this technique, the pores in the cell membrane are opened by introducing short high-voltage pulses, so that dUTP molecules can enter the nucleus [57]. These molecules consist of a fluorescent label and a nucleotide analog. During replication, early initiation sites are labeled around the regions where replication begins (Fig. 1.9b). However, the labeled nucleotides become rapidly depleted from the cell nucleus, preventing incorporation at later initiation sites. The ORM data is both single-molecule and high throughput which overcomes the shortcomings of both replication timing profile datasets and DNA combing data. However, the labeling is sparse (Fig. 1.9). DNA combing data illuminates whole replication domains, whereas these domains are not clear with ORM data. Since the fluorescent dyes are linked to the nucleotide analogues before adding them to the cell, the label concentration will be much smaller in ORM due to the larger size of fluorescent dyes compared to the nucleotide analogues. We will discuses the reasons for different labeling choices in the next paragraph.

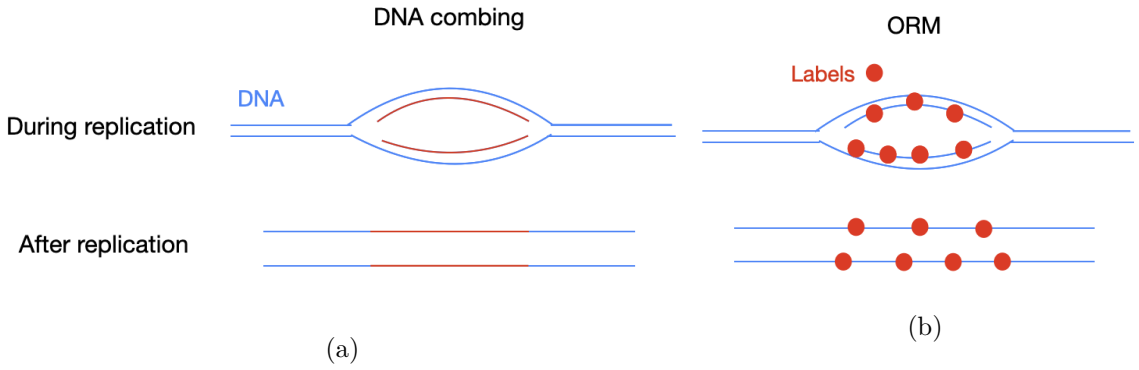


Figure 1.9: **Labeling in DNA combing and ORM experiments.** (a) Efficient labeling in DNA combing experiments. (b) Sparse labeling in ORM experiment. The labels which are nucleotide analogues attached to fluorescent dyes, incorporate into free DNA strands at replication forks.

After replication is complete, the chromosomes are fragmented by mechanical shearing and transferred into nano-channels for analysis [58, 55] (see Fig. 1.10) [59]. The nano-channels used in the ORM experiment have a square shape with a width of approximately

45 nm. 12000 channels can be contained in a nanofluidic chip of length 0.4 mm. These nanoscale-sized channels are specifically engineered for various scientific applications. Inside these channels, DNA fibres are linearized and prepared for imaging. The design ensures that the bending energy of DNA fibres inside the channels far exceeds the free energy of the system. The persistence length of a polymer, a measure of how far it can bend or twist before thermal fluctuations significantly affect its conformation, is an essential characteristic. For double-stranded DNA, this length is about 50 nm, approximately the same as the lateral size of the nano-channels. Consequently, DNA fibres are unlikely to bend significantly inside the channels because of their inherent rigidity. However, denaturing DNA fibres decreases their persistence length significantly resulting in folding of single-stranded DNA. Therefore, fluorescent dyes must be added to the nucleotide analogues before the DNA linearization as discussed in the previous paragraph.

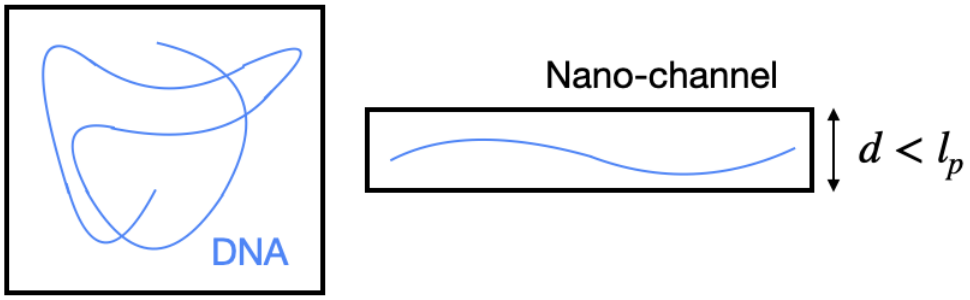


Figure 1.10: **DNA linearization inside a nano-channel.** The small diameter of the nano-channel prevents DNA molecules from bending.

High-resolution microscopy is used to observe and capture images of the DNA fibres within the nano-channels (Fig. 1.11). By analyzing the distribution of signal intensity from the fluorophores, the label positions along the DNA molecules were determined with 1 kb resolution. The green labels in Fig. 1.11 are restriction sites which bind to specific sequences along the genome. They are like barcodes that can be used to determine the position of the fibre along the genome. This approach enables the mapping of replication dynamics at a single-molecule level, providing valuable insights into the intricate processes of DNA replication across the entire human genome. One can then capture detailed information about the replication dynamics at a single-molecule level across the entire human genome.

ORM enables efficient linearization of several DNA fibres within the nano-channels [55]. DNA fibres are transferred into the nano-channel arrays in parallel with a short separation between the channels. This allows for storing a large number of fibres in a small enough space that the fibres can be visualized using a small number of fluorescent images. This leads to a higher coverage of the human genome (approximately 1000x coverage) than

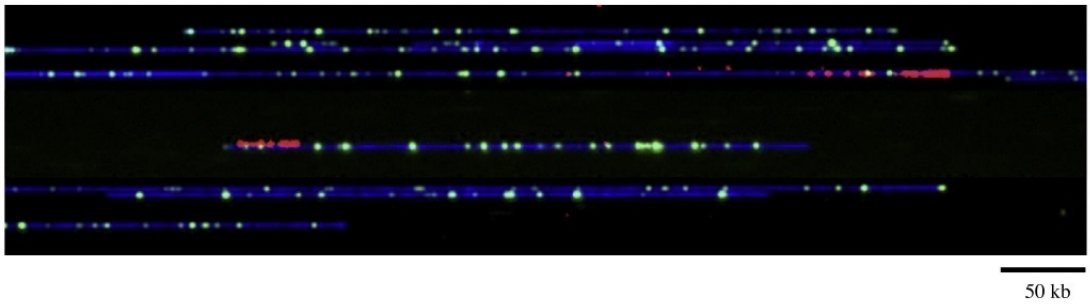


Figure 1.11: **Optical Replication Mapping image.** False-colour ORM image showing DNA fibres (blue), restriction-enzyme sites (green), and dUTP labels (red) in the nanochannels (not visible) [10].

traditional DNA combing methods, where many images are needed to provide 1x coverage of the genome.

However, for all their advantages, a new challenge has arisen in ORM experiments, which is the "sparse-labeling problem." Unlike DNA combing, where fluorescent dyes are added after linearization, ORM incorporates fluorescent dyes attached to nucleotide analogs prior to replication. This approach is necessary because denaturing the DNA fibres decreases their persistence length, making linearization difficult inside the nano-channels. However, the number of labels that can be incubated inside the nucleus is limited, reducing the incorporation rate of labels.

The prior analysis of ORM data has been limited to label density averaged over all the fibres [10]. This led to identification of initiation zones from peaks of label density. However, we will investigate the possibility of inferring initiation positions from each labeled single-fibre. Single-fibre inference of initiation events can potentially provide more information than population analysis of ORM data. More specifically, identifying each individual initiation events within the initiation zones provides information on the distribution of initiations inside the initiation zones and whether the initiation positions are sequence specific or they are spatially stochastic along the genome.

In this approach, we face the challenge of sparse labeling, which underscores the importance of understanding label incorporation for more precise initiation probability inference. Consequently, quantifying label incorporation in regions around initiation sites can aid inference methods, including those employing Bayesian inference to estimate initiation probabilities for individual fibres. In Section 1.3, we will present the mathematical frameworks for modeling replication kinetics and outline how these frameworks apply to model ORM experiments.

1.3 Modelling the replication kinetics

As discussed in Section 1.1, the DNA replication process in eukaryotes is complex, involving several molecular mechanisms, from licensing during the late M and G1 phases to activation

of potential initiations to unwinding of replication forks by helicases to the elongation phase [19, 17]. This process shows stochastic variation in the spatiotemporal organization of initiations and fork velocity [5, 6, 10]. Despite its complexity and stochastic nature, the average replication time is accurately regulated.

To study, understand, and quantify this intricate process, models have been developed at various scales. Here, we discuss the most common mathematical framework used to quantify replication kinetics in Subsection 1.3.1 and briefly review some of the inference approaches applied to the experimental data to infer replication program in Subsection 1.3.2.

1.3.1 One-dimensional Kolmogorov-Johnson-Mehl-Avrami (KJMA) model

Several mathematical frameworks have been developed to model replication kinetics in simple eukaryotes such as brewer's yeast (*S. cerevisiae*) [60, 26], fission yeast (*Schizosaccharomyces pombe*) [61, 62], and *Xenopus laevis* egg extracts [62, 63]. One widely adopted model, which has proven successful in both simple and more complex eukaryotic systems, is derived from the one-dimensional Kolmogorov-Johnson-Mehl-Avrami (KJMA) framework [64, 65, 66].

The KJMA model, initially developed in the 1930s to explain the stochastic kinetics of first-order phase transitions from liquid metals to the solid state, comprises three stages: (1) nucleation of solid domains, where solid islands begin to form at different points within the system; (2) growth, during which these solid islands increase in size; (3) coalescence, where solid islands merge and form larger islands until replication is completed (See Fig. 1.12 for a two-dimensional illustration). This model allows for deriving mathematical properties of phase-transition kinetics, such as the fraction of replicated regions as a function of time.

In the 1980s, K. Sekimoto analyzed the mathematical properties of the one-dimensional KJMA model, deriving a more detailed mathematical description of this model such as nucleation-domain statistics [67]. We conceptualize replication as a one-dimensional KJMA model, where chromosomes are depicted as one-dimensional lines [68, 69]. Each position along these lines represents a point along the chromosomes (see Fig. 1.13). In this representation, initiations are analogous to nucleations. The growth of replication bubbles due to the progression of two replication forks away from the initiation position mirrors the growth of islands in the KJMA model. Furthermore, the merging of replication bubbles corresponds to the coalescence stage in the KJMA framework.

According to this model, the probability of an initiation event at an unreplicated position $(x, x+\Delta x)$ along the chromosome and time $(t, t+\Delta t)$ after the onset of S phase is represented by an initiation rate $I(x,t)\Delta x\Delta t$. This model assumes there is no correlation between the initiation events. The correlation between origins has been studied in *X. laevis* [70] and human-genome initiation zones [10] showing no correlation outside of the scope of KJMA model. Additionally, the fork velocity is denoted as $v(x,t)$. These two factors intricately govern the replication kinetics. The fork velocity can be influenced by various aspects of the DNA replication process, such as fork stalling, where replication forks momentarily halt

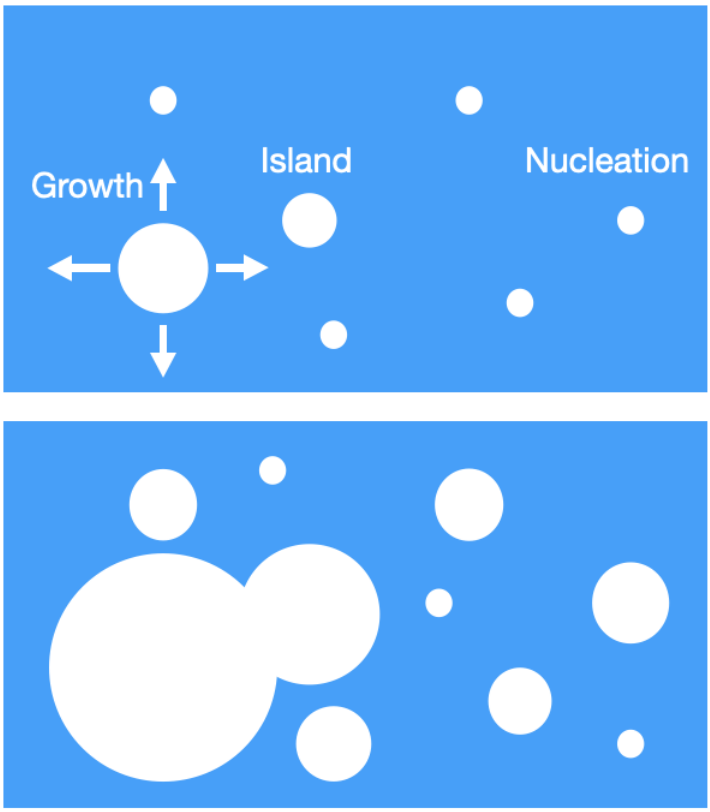


Figure 1.12: **Kolmogorov-Johonson-Mehl-Avarami model in two dimensions.** The white dots represent nucleation points, while the white disks signify solid domains or islands. The blue section represents the area where the phase transition has not occurred yet. The islands grow and subsequently merge, as the overall phase transition progresses.

before resuming movement [71]. Variations in fork velocity have been studied in budding yeast [72], and experiments have explored the fork velocity and inter-origin distance [73]. In this thesis, we assume that all forks travel at a constant velocity v. This assumption is based on previous research indicating that fork velocity remains relatively constant in budding yeast [7] and human cell lines [74].

This approach was used to describe DNA combing data in *X. laevis* early embryos [75], using a uniform initiation rate over genome that increased with time. Fitting models of *S. cerevisiae* to replication timing profiles based on microarray data shows that initiation sites are well characterized and sequence specific [26] and the total initiation rates (initiation rate integrated over the genome $I(t)$) of the initiation sites are bell shaped. Since embryonic cells divide more rapidly than somatic cells, a larger number of initiation events occur over the embryonic genome and leads to a more homogeneous distribution of origins.

The one-dimensional KJMA model has also been employed to simulate the human replication process, assuming N rate-limiting factors that bind to specific genome sites and initiate replication, with each factor representing one replication fork [74]. The initiation

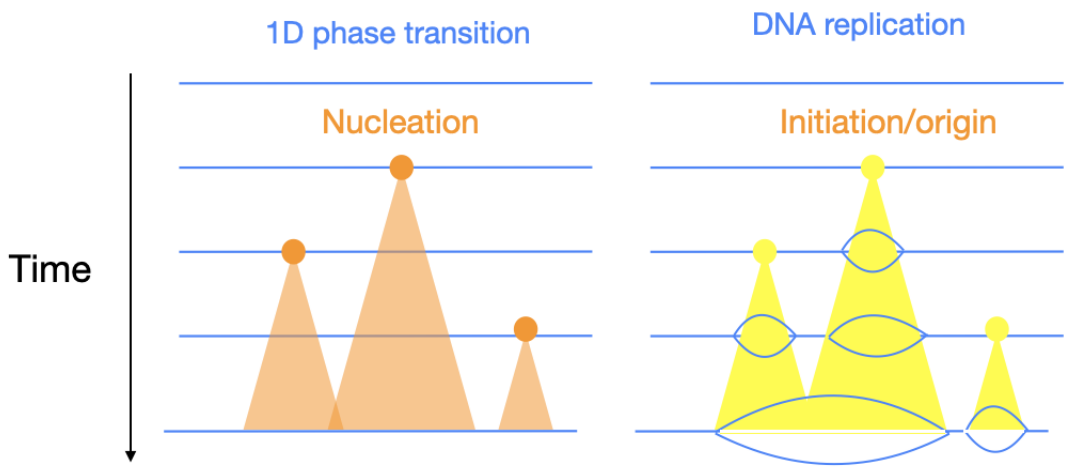


Figure 1.13: **One-dimensional KJMA for modelling the replication kinetics.** Analogy between one-dimensional KJMA elements (left) and replication kinetics (right).

rate in this model is expressed as a product of spatial and temporal components: $I(x,t) = f(x)g(t)$. Here, $f(x)$ is the "initiation probability landscape" (IPLS), representing the spatial component, while the temporal component is contingent on the probability of a single rate-limiting factor binding to the genome multiplied by the total number of available rate-limiting factors. This model has demonstrated successful applications to experimental data of replication timing profile with a resolution of 500 base pairs [76]. We will use this approach to simulate DNA replication in the human genome to generate simulated data that can be compared with experimental ORM data (see Section 2.1). The initiation probability landscape identified by this model correlates with DNase I hypersensitive sites (DHS). There is also evidence for correlation between the boundaries of Topologically Associated Domains (TADs) and replication domain boundaries [43].

Despite the inherent stochasticity of eukaryotic replication kinetics, embryonic cells accurately regulate the duration of the S phase, with only minor variations in duration [77]. The replication time is determined by the time of last coalescence. Simulations show that for a time-independent initiation rate, the variation in replication time is larger than measured in experiments. Researchers have shown a time-dependent initiation rate function where initiation rate increases in time, decreases the probability of having large gaps between origins; thus, the total replication time will be regulated. This also means that the total initiation rate will be bell-shaped because the available DNA content will decrease as the replication process progresses. Experimental data support a universal, bell-shaped temporal program of initiation [78]. There are various approaches to explain the bell-shaped temporal behavior of total initiation rate [79, 80, 78].

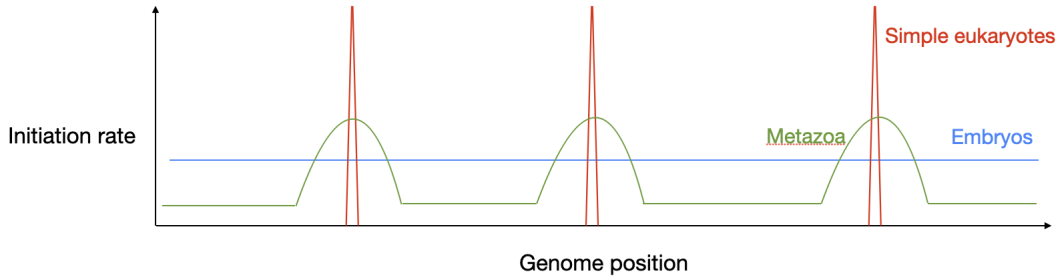


Figure 1.14: **Initiation rate of different organisms over the genome.** Embryonic replication kinetics can be modelled by a homogeneous initiation rate. In simple eukaryotes, the initiation events occur at specific positions along the genome. In metazoan genomes such as human genome, the initiation rate is a combination of homogeneous and localized initiation rate. The regions with higher probability to initiate are identified as initiation zones.

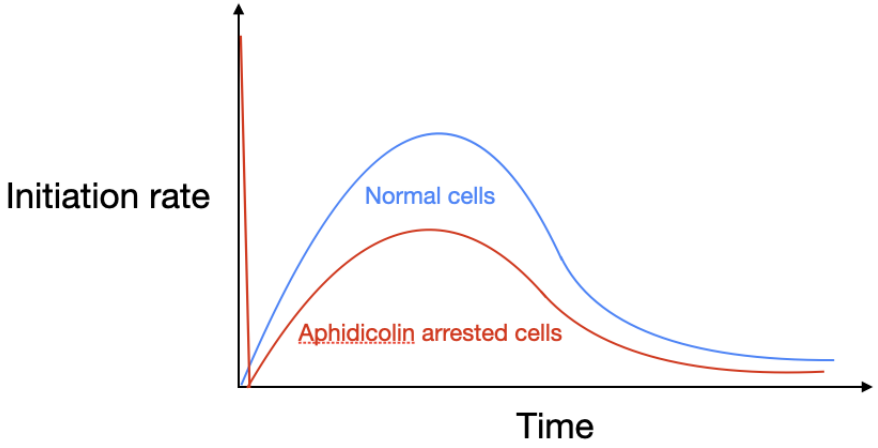


Figure 1.15: **Universal bell-shaped initiation rate.** Replication initiation rate in wild-type cells follows a universal bell-shape [78]. When cell's are arrested at the beginning of S phase by aphidicolin blocks, multiple initiation events occur after release of the aphidicolin blocks adding an effective delta-function initiation rate at $t = 0$ (red).

It is worth noting that recent research has detected 10–20% oscillations in fork velocity in bacteria changing as a function of temperature [81]. This study shows correlations between fork-velocity variation and mutation rate in *E. coli*. Studies modeling the effect of stochastic fork stalls in eukaryotes concluded that the effective fork velocity remains largely unaffected until the stall density surpasses a significantly higher threshold than observed in regular replication realizations [82]. Additionally, prior experiments and simulations of the eukaryotic DNA replication process indicate that assuming a constant fork velocity is a satisfactory approximation when examining resolutions greater than 1 kbp [7, 74]. Thus, we assume a constant fork velocity in this study.

To generate DNA replication kinetics, one needs to accurately infer quantities such as $I(x,t)$ and v using KJMA model. In the next section, we present previous approaches used to infer the replication program from available experimental data.

1.3.2 Inference of replication program from experimental data

The intricate and stochastic nature of biological and chemical processes within the small scale of a cell's nucleus pose several challenges in accurately identifying initiation sites and replication forks during the replication processes. Therefore, inferring replication realizations from datasets is crucial for understanding the DNA replication process. Various experimental approaches, ranging from bulk methods measuring average replication profiles to single-molecule techniques, have been used to infer the replication program in higher eukaryotes. In Section 1.2, we reviewed these experimental approaches. There are several experimental and modelling approaches to quantify replication kinetics. Here, we will highlight the inference techniques developed to extract replication-program information using the KJMA framework.

In these studies, the replication process is simulated using parameters that determine the initiation rate and fork velocity. By fitting to experimental data, which could either be DNA combing data or replication timing profiles, one can infer the initiation rate and fork velocity. Such an approach has been used to study *X. laevis* [75], *S. cerevisiae* [26], and human cells [74]. Studies have also investigated the impact of limited fibre length on the inference of replication kinetics from DNA combing data [83]. There are also studies on SMARD datasets, where two labels with different colours are inserted into the nucleus in order. The two-colour labeled data aids identification of replication fork direction [7].

Analytical approaches have also been used for inferring replication kinetics. The rate-equation approach uses inhomogeneous initiation rate and fork velocities to fit to right- and left-moving fork densities deduced from the SMARD data [7]. This method directly gives the mean-field kinetics of replicating DNA. These solutions are equivalent to simulations in the limit where an infinite number of simulations is performed. Moreover, the initiation rate can directly be deduced from the replication fraction [84]. Another analytical approach uses Gaussian process regression for inference [85]. This technique models the data as a Gaussian noise of the unreplicated fraction. The method uses Bayesian approach to infer initiation rate without making detailed assumptions about its functional form in the way required for curve-fit methods. Yet another Bayesian method uses Okazaki fragment (OF) sequencing data to infer local information about origin-firing activity [86]. Measuring mean replication timing (MRT) and replication fork directionality (RFD) is another way to infer the replication kinetics [85]. Based on these datasets, an inference approach has been developed using neural network for kinetic modelling of human genome replication [87].

In this thesis, we focus on Optical Replication Mapping data. This data comes from labeled DNA fibres, as with DNA combing. ORM has both advantages and shortcomings:

The ORM data is high throughput, but the labeling is sparse. In other words, DNA combing data illuminates whole replication domains, whereas these domains are not clear with ORM data (see Fig. 1.9). Therefore, the previous approaches used to infer DNA combing experiments are not sufficient for ORM data. To extract meaningful information from ORM datasets, Wang et al. [10] developed a model to describe the incorporation rate of labels around initiation sites based on an exponential-decay model for the number of labels inside the nucleus [10]. This model gave estimates of the inter-label distance distribution, initiation event labeling efficiency, and the correlation of labeling between IZs. However, we will see that this model has limitations when applied to other aspects of the ORM data.

To address these shortcomings, I introduce a more refined and generalized model to describe the incorporation patterns of labels in ORM experiments. The subsequent chapters will present in detail the limitations of the previous model, introduce the refined model, and explore how simulations were used to enhance the understanding of ORM data. These efforts aim to improve the accuracy and reliability of DNA replication studies using ORM technology.

Chapter 2

Numerical methods

Optical Replication Mapping has produced a novel high-throughput, single-molecule dataset for studying DNA replication kinetics in the human genome [10]. However, the sparse-labeling problem decreases the resolution of initiation position inference. Consequently, understanding the label distributions around the initiation events can provide more insight about the position of the initiation events. To achieve this, we simulated early replication kinetics and fluorescent label incorporation, based on our models (as detailed in Sec.2.1). Then, we fit the outcomes of stochastic simulations to the experimental data (as outlined in Sec.2.2) to identify the optimal model for the Optical Replication Mapping experiment.

2.1 Simulation

Simulation of the Optical Replication Mapping (ORM) experimental data has two key stages: [68] (1) Simulating the early replication dynamics and identifying the replication tracks along the DNA fibres. (2) Generating the labels incorporated into the DNA fibres using the incorporation rate and the identified replication tracks.

2.1.1 Simulation algorithm

The replication kinetics of an organism are influenced by the properties of origins, including their licensed positions and ability to initiate throughout the S phase, as well as by the properties of fork progression [68]. As discussed in Sec. 1.3, we employ one-dimensional KJMA processes to model replication kinetics. Various algorithms have been developed for simulating this process, including Ising-model-like algorithms [75] and Double-list algorithms [88]. In this study, we use the Phantom-nuclei algorithm, which is faster than the other two approaches [68].

The phantom-nuclei algorithm consists of two stages (see Fig. 2.1):

(i) We generate potential initiation sites in the two-dimensional space-time plane using inhomogeneous Poisson processes with an average rate of $I(x,t)\Delta x\Delta t$ at (x,t). The position and initiation times of every potential initiation site are stored in two vectors.

(ii) For each potential initiation site (indexed by i), with position x_i and nominal initiation time t_i, one calculates the position of the replication tracks at the reference time point t by $r_i = x_i + v(t - t_i)$, where r_i indicates the position of a potential replication fork coming from initiation event i and at time t. Finally, we find the minimum timing of the replication tracks coming from all the potential initiation sites to find the "real" replication track.

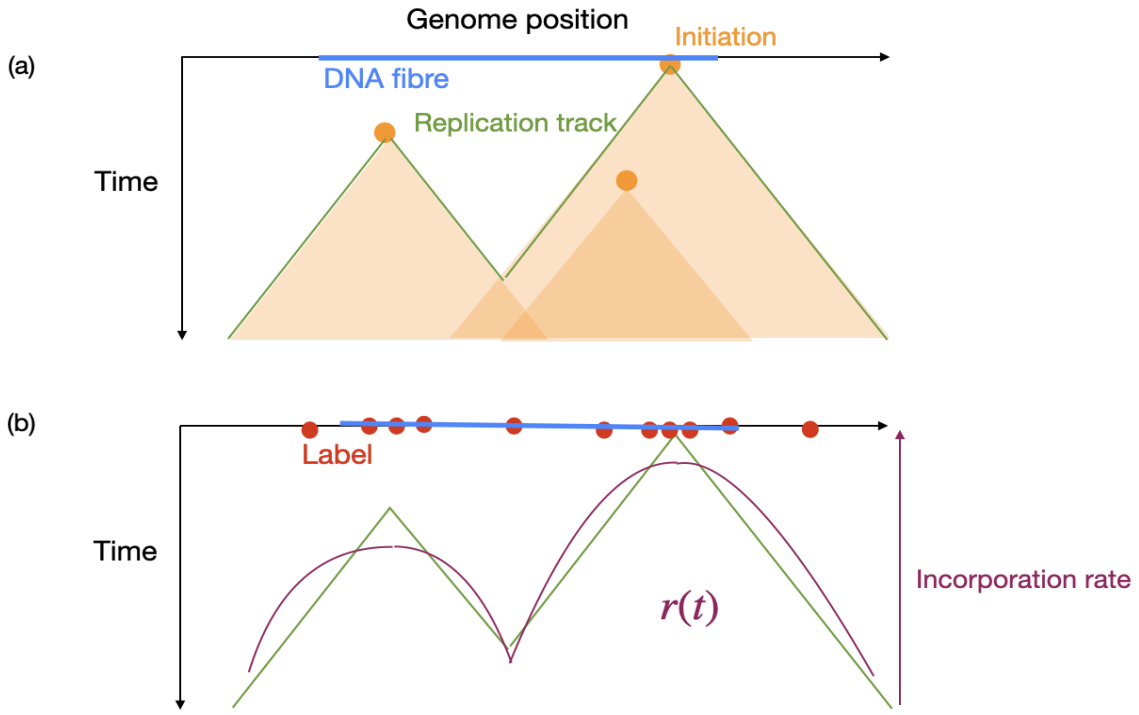


Figure 2.1: **Phantom-nuclei algorithm.** [68] (a) Initiation events occur at different points in space-time (orange dots). The potential replication track coming from each initiation event is calculated (orange triangles). The true track of replication forks is calculated by finding the minimum of all the potential tracks (green lines). We simulate the region around a DNA fibre. (b) The incorporation rate (purple line) is calculated by applying $r(t)$ to the replication track timing. The shorter the time after start of replication, the higher the chance that labels incorporate into available sites. Label positions are generated by using the incorporation rate as the probability function (red markers). Some of the labels are incorporated into the DNA fibre, and some lie on DNA that is beyond the edge of the fibre.

To generate the positions of labels, we employ the incorporation rate $r(t)$, representing the probability of a label incorporating at each position along the genome (refer to Fig. 2.1). Importantly, this quantity does not vary with position, since label density remains homogeneous throughout the cell's nucleus. A detailed explanation of the incorporation-rate calculation is provided in Chapter 3.

Simulation of replication kinetics is carried out individually for each fibre and the surrounding regions. The fibre length is taken randomly from the fibre length distribution(see Fig. 3.1a). This approach is taken because replication forks originating from distant regions will not reach the fibres within the simulation's time limit. The time limit for the simulation (set at 90 minutes) is chosen to ensure that $r(t)$ is sufficiently small after the cutoff, preventing incorporation from occurring.

The incorporation of labels at position x is determined by running Poisson processes with an average rate of $r(t(x))\Delta x$ over the fibres, where $t(x)$ represents the time at which a replication fork reaches position x along the fibre (see Fig. 2.1). After the simulation is completed, three plots are extracted from the data. Inter-label distance distribution, distribution of the number of labels on the fibres, and the average number of labels as a function of fibre-length (see Fig. 3.2). More details are discussed on how to generate these plots in Section 3.1. These plots are then used to fit to the experimental data using methods discussed in Section 2.2. The fit parameters are the parameters determining the initiation rate $I(x,t)$ and label incorporation rate $r(t)$. More details about the fit parameters is discussed in Sections 3.2 and 3.3. The loss function used for fitting is sum of chi-squared ($\chi^2 = \sum_i \left(\frac{o_i - e_i}{\sigma_i}\right)^2$) of all the plots. Here o is the observed value, e is the experimental value, and σ is the error for the experimental values. The errors are calculated assuming the experimental values have Gaussian distributions.

2.1.2 Testing the simulation

Before exploring whether the model predictions fit the experimental data, we need to test the simulations. In this subsection, we discuss a realization where we compare the simulated ORM data with analytical solution.

To do this, we assume the rate of initiation is uniform along the genome and the initiation events are localized at the start of replication process $I(x,t) = 2I_0\delta(t)$, where I_0 is the rate of initiation over each position along the genome and $\delta(t)$ is the Dirac delta function. However, for the analytical approach, we use the replication track of the closest initiation event to the DNA fibre; taking multiple initiations would complicate the calculations. The single-initiation assumption works for small initiation rates where there is no initiation for most of the simulated fibres. Then, since the total number of initiations follows a Poisson distribution, the number of replication realizations with more than one initiation will be small enough to ignore in our calculations. Therefore, the probability of observing n labels on a DNA fibre with length L, denoted $P_L(n)$ will be, will be

$$P_L(n) = \int_{-\infty}^{\infty} \mathrm{d}x P_i(x) P_L(n|x), \tag{2.1}$$

where $P_i(x)$ is the probability of observing the closest initiation event to the center of DNA fibre at position x along the genome. Also, $P_L(n|x)$ is the probability of observing n labels

coming from an initiation at position x. For the former probability function, we have

$$P_i(x) = I_0 \mathrm{e}^{-2I_0 x}, \tag{2.2}$$

where I_0 is the probability of observing an initiation event at x and $\mathrm{e}^{-2I_0 x}$ is the probability of not having an initiation event in the $(-x, x)$ interval. For the probability to observe n labels, we have

$$P_L(n|x) = \text{Poisson}\left(n, \int_{-L/2}^{L/2} \mathrm{d}x' r(x - x')\right), \tag{2.3}$$

where $r(x)$ is the label incorporation rate giving the probability of a label incorporating to the genome in distance x from an initiation event. As discussed in section 1.2 the labels are made of thymine analogues incorporating to adenine. Since the resolution of label positions is 1 kb, I assume there is no sequence dependence of label incorporation. We use an exponential decay function for the incorporation rate [10].

$$r(x) = r_0 \mathrm{e}^{-x/l}, \tag{2.4}$$

where r_0 is the incorporation constant and l is the depletion length. We will discuss the calculation of incorporation rate in Chapter 3.

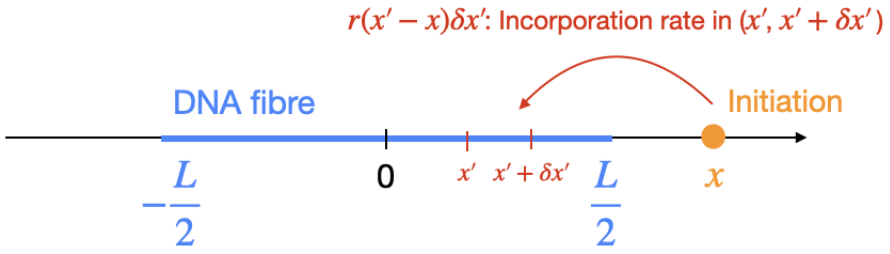


Figure 2.2: **Model of label incorporation.** The closest initiation is at position x; there is no other initiation in $(-x, x)$. The label incorporation coming from this initiation event is calculated using the incorporation rate function.

To find the number of labels, we average over all initiation realizations. By inserting Eq. 2.4 into Eq. 2.3, we find

$$P_L(n) = \int_{-\infty}^{\infty} \mathrm{d}x I_0 \mathrm{e}^{-2I_0 x} \text{Poisson}\left(n, \int_{-L/2}^{L/2} \mathrm{d}x' r_0 \mathrm{e}^{-\frac{|x'-x|}{l}}\right). \tag{2.5}$$

The final label number distribution is calculated by numerical integration of Eq. 2.5.

We compare the simulations and analytical calculations in two regimes: one where the average number of initiations is ≈ 0.1, and one where average number of initiations is ≈ 1,

as shown in Fig 2.3. We expect that the simulation and the analytical approximation align for the former case because the number of instances with multiple initiation is extremely small. However, for the latter case, we expect to see deviations between simulation and analytical solutions, where the fraction of fibres with higher number of labels is higher for the simulations because multiple initiations lead to more labels. Indeed, we see this pattern in Fig. 2.3 for different sets of parameters.

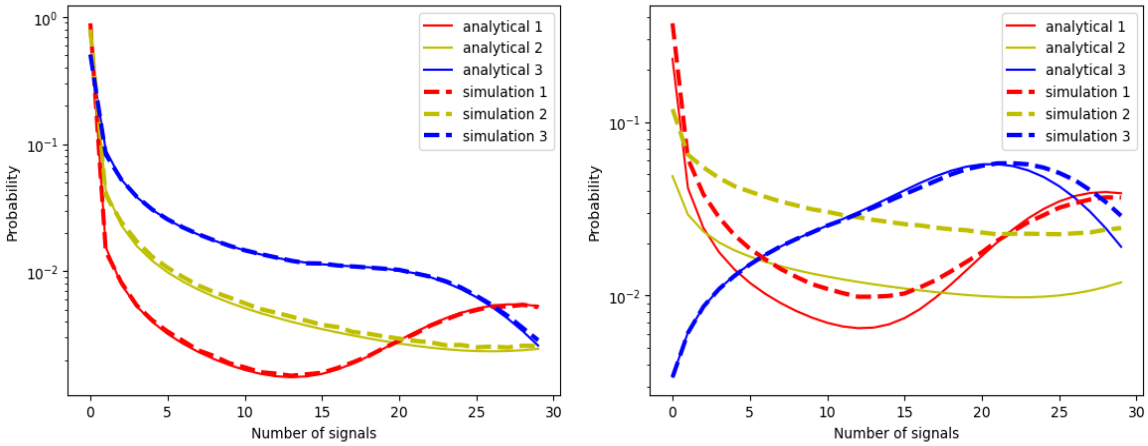


Figure 2.3: **Comparing simulation and analytical number of label distribution.** (a) $I_0 L \approx 0.1$ where either there is usually no or a single initiation event along the DNA fibre. (b) $I_0 L \approx 1$ where there are many realizations with multiple initiations. Three different sets of parameters for incorporation rate function (see Eq. 2.4) are used for both (a) and (b). Parameter set 1: $r_0 = 0.5$ kb^{-1}, $l = 30$ kb Parameter set 2: $r_0 = 0.3$ kb^{-1}, $l = 100$ kb Parameter set 3: $r_0 = 0.1$ kb^{-1}, $l = 300$ kb. I_0 and L are fixed.

In this section, we showed how we simulate the ORM experiment and tested our algorithm. In order to use this as a tool to understand the ORM data, we need to fit the simulated data to the experimental data. This brings its own challenges, as discussed in Sec. 2.2 in detail.

2.2 Data fitting

To test our models, we need to fit the results of simulations to the data to find the optimal parameters. The challenge is both the stochasticity of the models and the long run time of the simulations. (To simulate a million fibres takes about 10 s). To overcome these challenges, we tried different approaches, including simulated annealing, stochastic gradient descent, and grid search to find the optimum parameters for our models. We will discuss the datasets we use for model selection and the fitting approaches in this section.

2.2.1 Simulated annealing

Simulated annealing is a probabilistic optimization algorithm inspired by the annealing process in metallurgy [89]. Annealing involves heating a material to a high temperature and then gradually cooling it to remove defects and reach a low-energy internal structure. Similarly, in computational optimization, simulated annealing is used to find an approximate solution to an optimization problem by exploring the solution space and gradually reducing the probability of accepting worse solutions over time (Fig. 2.4).

Simulated annealing starts with an initial trial solution to the optimization problem. Simulated annealing uses a temperature parameter that controls the probability of accepting a worse solution. Initially, the temperature is set high, allowing the algorithm to accept worse solutions with higher probability. The temperature is then gradually reduced over time, according to a predefined schedule. The algorithm generates a neighbouring solution by making a small change to the current one. The neighbourhood exploration is a key aspect of the algorithm, and it defines how the search space is traversed. If, after taking a step towards a neighbouring position, the optimization function decreases, the step is accepted. If not, the probability of taking the step is $e^{-\Delta E/T}$ as with the Metropolis algorithm, where T is the temperature and ΔE is the increase of the optimization parameter. Then the temperature is updated according to the predefined schedule, typically reducing it over time.

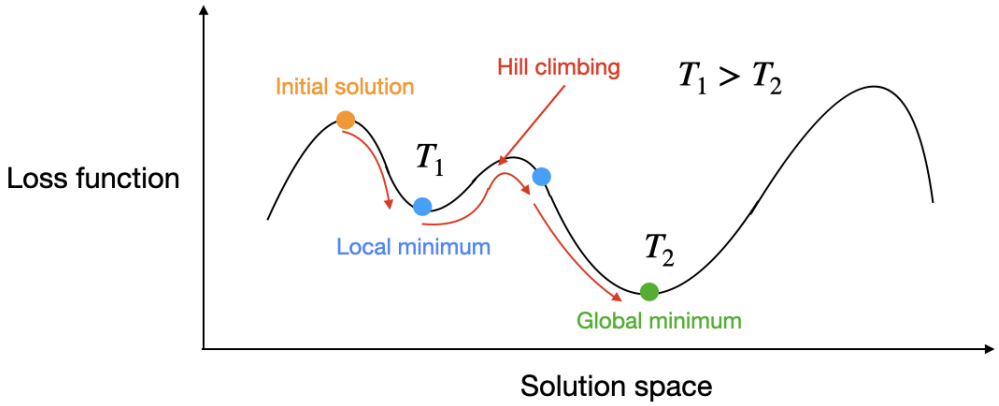


Figure 2.4: **Simulated annealing algorithm.** This figure shows how simulated annealing finds the minimum loss function. The algorithm finds local minima but it can climb the hills of these regions given a probability related to system temperature find the global minimum.

The stochasticity of the loss function does not impact the final result since the algorithm does not use the gradient of the loss function for searching the local space.

We generated reproducible results using this method (see Chapter 3). This fit is done in Python using the "dual_annealing" function in SciPy [90]. Implementing Dual Annealing optimization [91]. Where the temperature is decreased approximately proportional to the

inverse of the simulation time step. The step size is given by a Cauchy-Lorentz distribution, allowing for taking occasional long steps enabling the search algorithm to escape local minima. To test the optimum solutions, we repeated our fits, and the final results were reproducible. However, a challenge with this approach was the long time needed for the fits.

2.2.2 Gradient descent

In an attempt to speed up the model fitting, we explored a stochastic gradient descent algorithm. Gradient descent is an optimization algorithm used to minimize a function by iteratively moving in the direction of steepest descent of the function's gradient [92]. It is widely employed in various machine-learning algorithms, linear regression, logistic regression, neural networks, and more [93].

The algorithm starts with an initial guess for the parameters of the function being optimized. At each iteration, the gradient of the function with respect to the parameters is computed. The parameters are updated by taking a step in the opposite direction of the gradient. The size of the step is determined by the learning rate, which is a hyperparameter that is chosen before the optimization process begins. These steps are repeated iteratively until a stopping criterion is met. This criterion can be a maximum number of iterations, reaching a predefined tolerance level, or other conditions specific to the problem being optimized. In order to find the minimum number of simulation realizations to decrease the runtime of optimization, we need to make sure the variance of the loss function is smaller than the change in the loss function [93].

Using these ideas, we ran an optimization algorithm based on gradient descent. The results, however, did not converge to the expected minimum. One possible reason is that the stochasticity of the loss function does not let the search algorithm to stop in a local minima. The stochasticity decreases with increasing the number of simulation realizations, but doing this increases the runtime. To get more intuition into the parameter landscape, we use the grid-search algorithm presented in the next subsection.

2.2.3 Grid search

Grid search is a way to explore the value of an optimization function in the parameter space. To better understand the aspects of the loss functions resulting from our models and test the results of our systematic optimization approach (simulated annealing), we have plotted the loss function as a function of aphidicolin-blocked initiations occurring at $t = 0$ and early initiation events occurring at $t > 0$ (see Fig. 2.5).

The contour plot of the loss function shows that the global minimum found using the simulated annealing does fall into the region with minimum loss function.

In this chapter, we tried three approaches to optimize the stochastic loss function resulting from simulation of ORM data. The most reliable approach that we found was simulated annealing plus checking the final results with a grid search to make sure the

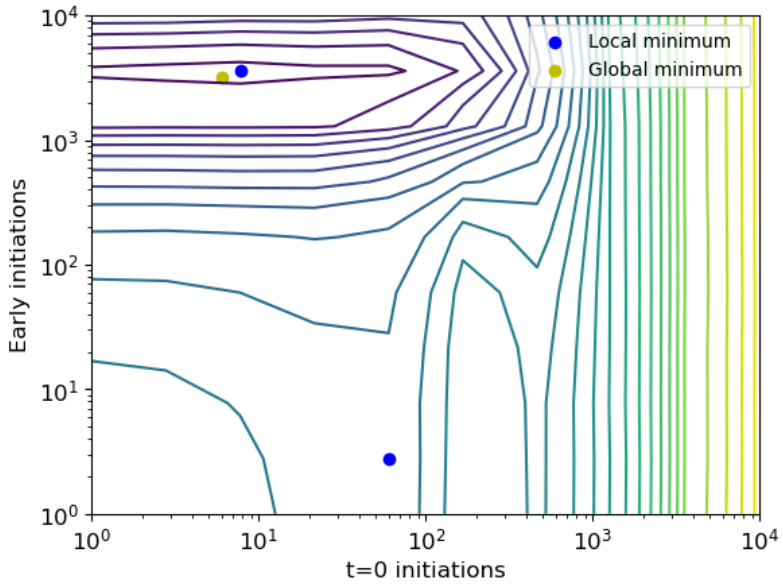


Figure 2.5: **Loss function vs initiation.** The loss function is calculated for two out of three parameters of the time dependent model, aphidicolin arrested initiations (y-axis) and early initiations (x-axis).

global minimum has been found according to the loss function landscape. The gradient-descent algorithm, however, has not been useful for finding either local or global minima, which seems to be the result of stochasticity of the loss-function. In the next chapter, we will discuss the results of fitting our models to the ORM data.

Chapter 3

Results

In Section 1.2, we discussed Optical Replication Mapping (ORM), a single-molecule technique used to investigate the distribution of initiation events over the human genome by mapping labeled DNA fibres that are stretched out in nano-channels [10]. This chapter presents the ORM data and highlights the label-incorporation modelling approach used in this study, as detailed in Section 3.1. Initially, a model was developed that assumed an exponentially decreasing number of labels within the nucleus. However, I found that this model failed to account for certain aspects of the ORM data. To address this limitation, I refined the model by incorporating simulations of the early replication process (Section 2.1) and fitted the new models to the experimental data (Section 2.2), as discussed in Section 3.2. I also introduced a combination of aphidicolin-blocked initiations ($t = 0$ initiations) and early initiation events into the model, discussed in Section 3.3. These refinements led to generalized models that improved the agreement with the experimental data. Despite these advances, these models, while promising, are still a work in progress and do not provide a complete description of ORM experiments.

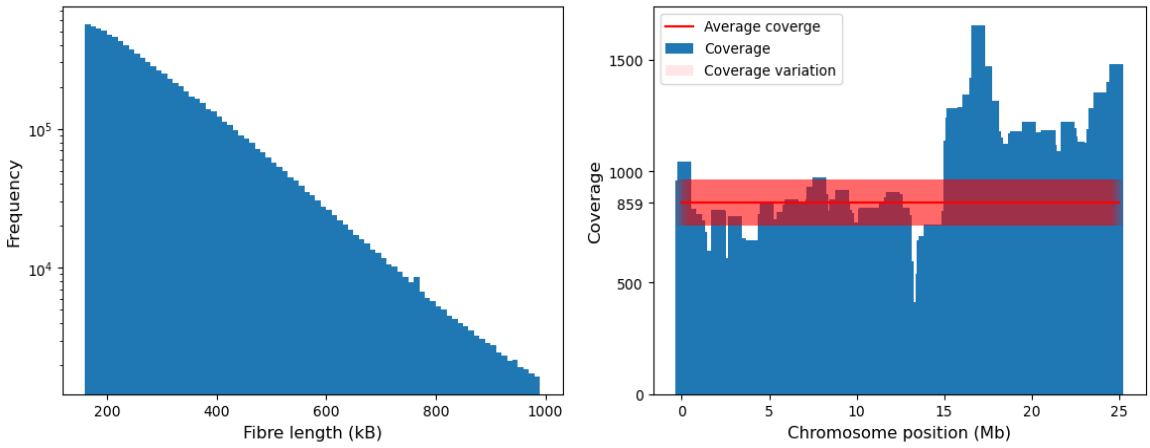
3.1 Previous analysis of Optical Replication Mapping (ORM) data

As discussed in Subsection 1.2.4, the sparse labelling of ORM data gives rise to a significant inference challenge. Therefore, understanding the label distribution around initiation events becomes essential to infer initiation position with good precision. In this section, we introduce the ORM data and also a model that was previously used to explain the incorporation probability of fluorescent labels [10].

3.1.1 ORM data

The ORM experiment is conducted for both synchronous and asynchronous cultures of cells. In synchronous cultures, label incubation can occur at various times during S phase (0, 5, 10, 20, 30, 45, 60, and 90 minutes). In this thesis, we focus on the 0-minute dataset.

This choice is strategic: at other times in S phase, many replication forks move away from initiation events, causing labels to incorporate far from the initiation positions. Therefore, these instances do not provide sufficient information for inferring initiation positions. The 0-minute dataset comprises approximately 10 million fibres ranging from 150 kb to 2.2 Mb in length, providing about a 1000-fold coverage of the human genome. The coverage is uniform, showing only slight variations of approximately 12% as shown in the Fig. 3.1b. The variation in the coverage comes from the limited fibre length of the fibres (as detailed in Fig. 3.1a), which shows that the average fibre covers approximately 300 kb of the genome.



(a) (b)

Figure 3.1: **ORM fibres and coverage.** (a) Fibre length distribution for 0-minute ORM data. (b) Fold-coverage of DNA fibres over Chromosome 1 (The number of fibres covering bins of size 1 Mb). The red shaded region shows one standard deviation about the average.

The estimate of fluorophore positions within each fibre define is made by modeling the signal intensity and the number of photons emitted from each fluorophore as a Poisson distribution [10]. This modeling approach allows us to determine the positions of fluorescent labels (red) and restriction sites (green) (Fig. 1.11). Using the label density, one can define and identify initiation zones. The label density is obtained by calculating the percentage of fibres containing ORM labels in 10 kb sliding windows with 1 kb steps. Then, through smoothing and peak identification, initiation zones are located. Approximately 5000 initiation zones were identified, containing around 20% of the labels [10]. The remaining 80% of the labels are distributed elsewhere at lower densities, including in late-replicating regions of the genome. This work also investigates for non-specific label incorporation and concludes that there is no significant background noise.

It is worth noting that this analysis does not explore the single-molecule aspects of the ORM data. To extract more information, such as investigating the distribution of initiation events inside the initiation zones, and correlation of neighbouring initiations, the authors in [10] developed an exponential-decay model for the label incorporation (Subsection. 3.1.2).

This model was employed to study the distribution of labels inside initiation zones, the accuracy of initiation position inference, and correlation of labelling in neighbouring initiation zones. However, we will show that this model does not agree with ORM data.

To study the ORM experiment, we focus on three plots that I created from the dataset: (i) the inter-label distance distribution, "where inter-label" defined as distance is the distance between two neighbouring labels; (ii) the distribution of total number of labels along the DNA fibres; and (iii) the average number of labels as function of fibre length (Fig. 3.2). We will use the data in these plots in order to fit to model predictions (see Section 3.2 and 3.3).

In the next subsection, we will discuss the previous model developed to explain the ORM data.

3.1.2 Modelling the inter-label distance distribution

As discussed in Subsection 1.3.2, modelling the label incorporation pattern in replication domains is necessary for inferring more detailed information about the initiation positions. A first approach assumes that the total number of labels decreases exponentially, with the incorporation of labels proportional to the number of labels that are inside the nucleus. Therefore,

$$r(d) = r_0 \mathrm{e}^{-d/l}, \tag{3.1}$$

where $r(d)$ represents the incorporation rate, the probability that a label incorporates into free nucleotides per genome length (kb^{-1}) at a distance d from the initiation event [10]. This model is based on some simplified assumptions: i) The decrease in the number of labels is proportional to the number of labels, indicating an exponential decay; ii) The probability of labels incorporating into free nucleotides is proportional to the number of labels; iii) The initiation events start at $t = 0$, on account of the aphidicolin-block, which freezes the replication forks. To assess the validity of this model, we tested it by fitting to the inter-label distance distribution. The probability for two labels to be at a distance d is

$$p_d(d) = \int_0^\infty \mathrm{d}x p(x, x+d) = \int_0^\infty \mathrm{d}x p(x) p(x+d) p_{\notin}(x, x+d), \tag{3.2}$$

where $p(x, x+d)$ represents the joint probability of finding one label at position x and another at $x+d$, without any label in between them. Since label incorporation is independent of genome position, $p(x, x+d)$ can be decomposed into independent probabilities of finding labels at x ($p(x)$), $x+d$ ($p(x+d)$), and the probability of not finding labels in between ($P_{\notin}(x, x+d)$). The probability to find labels is proportional to the incorporation rate. For the absence of labels in a given region, we have

$$P_{\notin}(x, x+d) = \int_0^\infty \mathrm{d}x \exp\left\{\int_x^{x+d} \mathrm{d}x_0 r(x_0)\right\} = \int_0^\infty \mathrm{d}x \exp\left\{\int_x^{x+d} \mathrm{d}x_0 r_0 \mathrm{e}^{-\frac{x_0}{l}}\right\}. \tag{3.3}$$

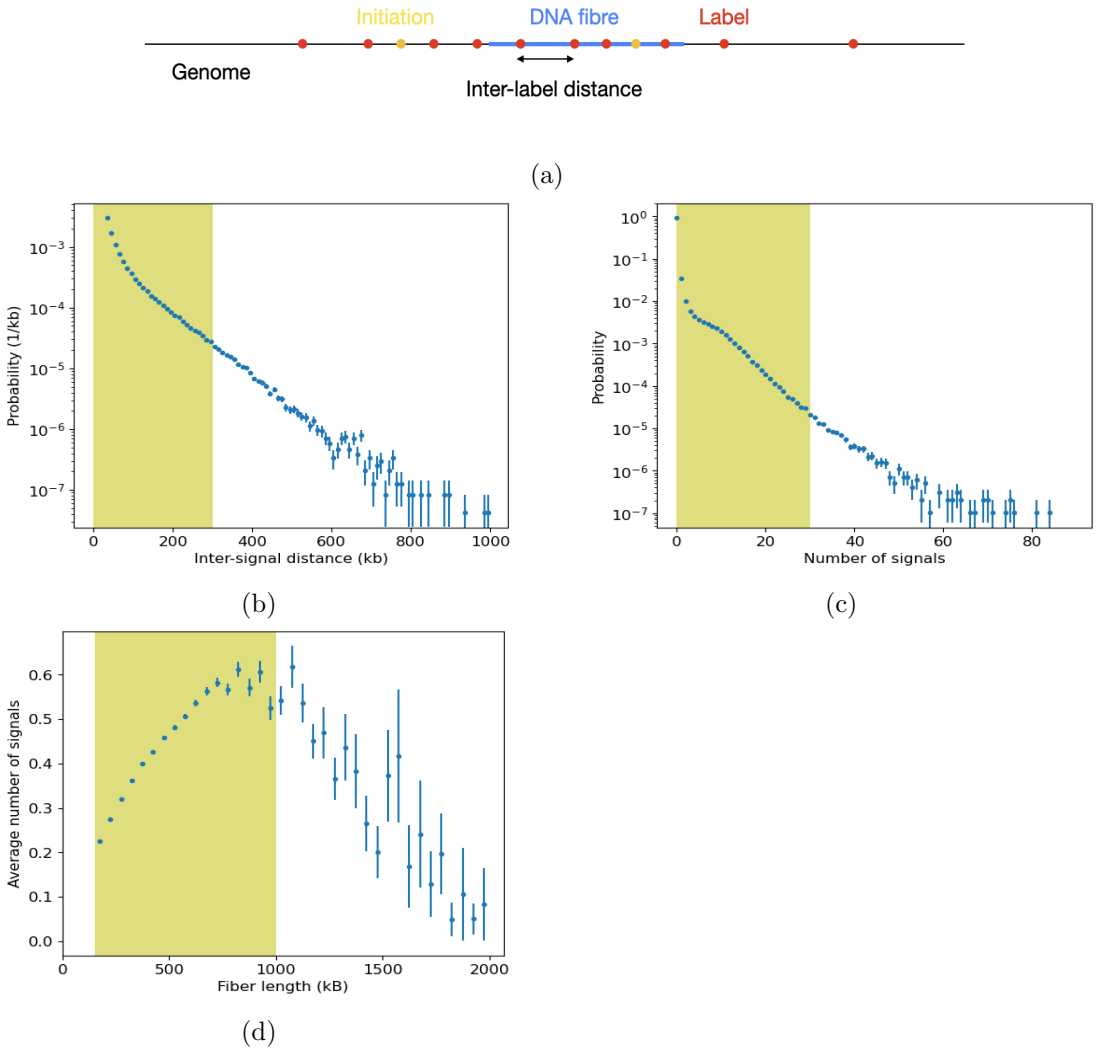


Figure 3.2: **Simulation of labeled fibres.** (a) The region around initiation events is labeled. Both labels and initiations can be either inside or outside of the fibres. The distance between two neighbouring labels is inter-label distance. By counting the number of labels on the fibre, we can deduce the distribution of number of fibres and average number of fibres as a function of fibre length. (b), (c), and (d) show the plots extracted from the ORM data. (We fit our models to the shaded region of the plots). (b) The inter-label distance distribution. (c) Distribution of number of labels. (d) Average label number as a function of fibre length.

Thus, by substituting Eq. 3.3 into Eq. 3.1, we find

$$p_d(d) \propto \int_0^\infty \mathrm{d}x \exp\left\{-\frac{2x+d}{l} - \int_x^{x+d} \mathrm{d}x_0 r_0 \mathrm{e}^{-x_0/l}\right\}, \tag{3.4}$$

Calculating the integral using Maple (https://www.maplesoft.com), we find

$$p_d(d) \propto \frac{\mathrm{e}^{r_0 l}(r_0 l \exp\left\{r_0 l \mathrm{e}^{-d/l}\right\}) + \mathrm{e}^{r_0 l + d/l} - (1 + r_0 l)\exp\left\{r_0 l \mathrm{e}^{-d/l} + d/l\right\})}{(\mathrm{e}^{d/l} - 1)^2}. \tag{3.5}$$

Equation 3.5 was successfully fit to the inter-label distance distribution [10] (Fig. 3.3). Consequently, this model was used to infer the initiation positions, estimate the efficiency of initiation zones, and analyze the distribution of labels within initiation zones.

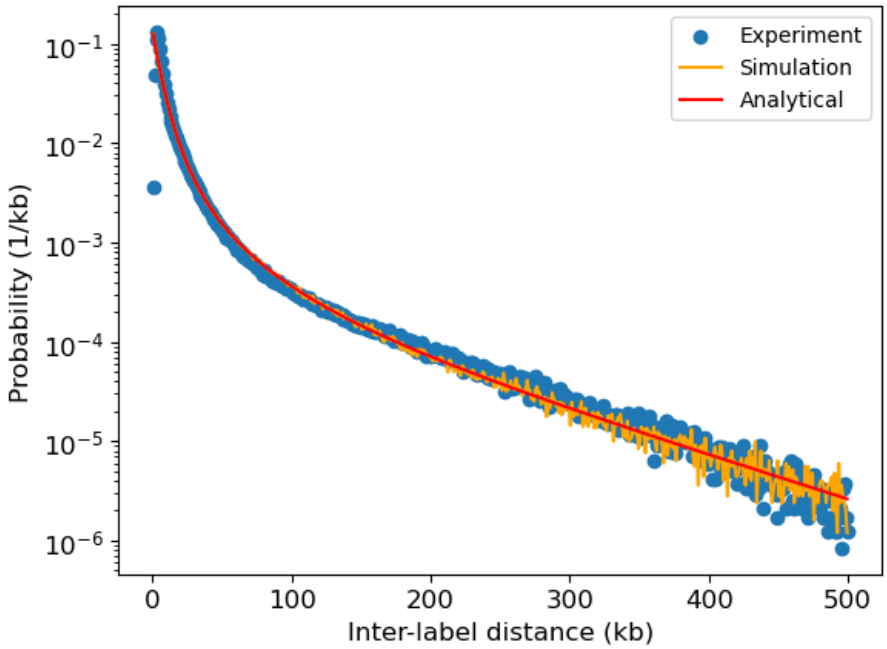


Figure 3.3: **Inter-label distance distribution.** The curve is the result of fitting Eq. 3.5 to the data ($l = 99 \pm 8$ kb, $r_0 = 0.26 \pm 0.04$ kb^{-1}). The orange curve shows the result of simulating the the inter-label distance distribution using the fit parameters.

However, by further analyzing this data, I found serious discrepancies with experiments. In Eq. 3.1, it is assumed that the fibre size is infinite, whereas the fibres analyzed in the experiment have a finite size (see Fig. 3.1). Additionally, the average number of labels on a fibre is

$$\bar{N} = 2\int_0^\infty \mathrm{d}x\; r(x) = 2\int_0^\infty \mathrm{d}x\; r_0 \mathrm{e}^{-x/l} = 2r_0 l = 53 \pm 4. \tag{3.6}$$

Here the upper limit of infinity in the integral corresponds to the infinite fibre-length assumed in the previous calculation of inter-label distance distribution (see Eq. 3.4). This number is much higher than the observed value of average number of labels on a fibre for fibres with at least one label (i.e., at least one initiation event occurring), which equals 4.12. To address the observed discrepancy, we modified this model and incorporated additional datasets for testing (Fig. 3.2). Recognizing the limitations of relying solely on the inter-label distance distribution, we extended our analysis to include two other types of data: the

distribution of label counts and the average label count as a function of fiber length. Unlike previous studies that assumed infinite fiber length when calculating inter-label distance distribution, our model accounts for the finite length of fibers. The infinite fibre-length assumption made in the previous analysis increases the number of labels on the fibres significantly, specially since the replication tracks can come from initiation outside of the fibres (see Fig. 3.2a). In the following section, we present this refined model and leverage all three datasets to enhance our understanding of the Optical Replication Mapping (ORM) experiment.

3.2 Limited-fibre length model

The initial approach to address the earlier discrepancies involves considering the observed length distribution. We used simulations (Section 2.1) to generate DNA fibres and incorporate labels into them based on the model. Initiation events can take place either inside or outside the fibres (Fig. 3.2). Using the replication track timing, we sample from Poisson processes to generate the labels on the fibres. The incorporation rate can be calculated easily by substituting time for distance ($d = vt$) in Eq. 3.1. Thus,

$$r(t) = r_0 \mathrm{e}^{-vt/l} = r_0 \mathrm{e}^{-t/\tau}, \tag{3.7}$$

where $\tau = l/v$ is the depletion time scale.

We applied our model to analyze three distinct datasets derived from the ORM data (c.f. Eq. 2.2): the inter-label distance distribution, the distribution of the total number of labels on each fibre, and the average number of labels as a function of fibre length (Fig. 3.4). The parameters governing these curves include the incorporation constant r_0, the depletion time τ, and the initiation rate constant I_0. Based on the prior analysis, incorporation rate parameters were determined as $l = 99 \pm 8$ kb and $r_0 = 0.26 \pm 0.04$ kb^{-1}. For fitting, we used the nominal values of these parameters. Notably, the sole fitting parameter in this context is I_0.

In Fig. 3.5a, the inter-label distance distribution shows a decrease in the tail, a result of limited fibre length that constrains inter-label distances. On the other hand, the total number of labels on each fibre exhibits a different pattern, with the slope of its tail significantly exceeding experimental data. When we consider the average number of labels as a function of fibre length, the model aligns with experimental data for shorter fibre lengths but deviates from the linear trend for longer fibre lengths.

Despite its limitations in explaining several aspects of the data, this model incorporates the assumption of limited fibre length. Notably, the average number of labels on the fibres ($\bar{N} \approx 0.32$) agrees with experimental observations, an improvement over the previous model. We also extend the fitting to all three parameters I_0, r_0, and l. This approach improved the fit slightly (Fig. 3.5); however, the simulated data remains far from experimental results.

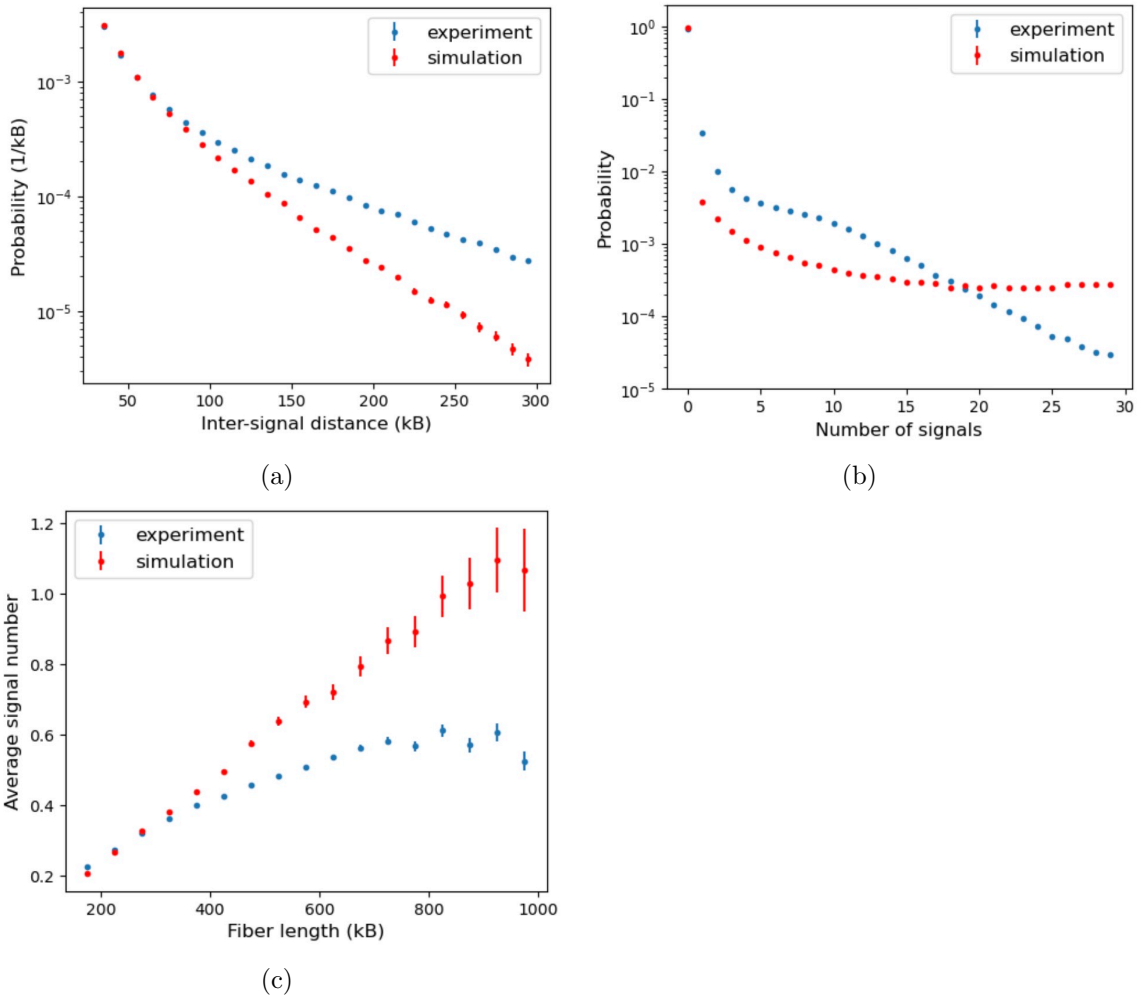


Figure 3.4: **Result of limited-fibre length model 1.** (Varying for I_0 while fixing $r_0 = 0.26\ \mathrm{kb}^{-1}$ and $\tau = 61$ min). (a) Inter-label distance distribution. (b) Distribution of number of labels. (c) Average label number as a function of fibre length.

One potential explanation for the disparities between experiments and simulated data is to take into account the initiations occurring after $t = 0$, which reduces the likelihood of label incorporation. This might eliminate the local minima in the number of the label distribution and increase the probability of observing labels at longer distances, thereby explaining the tail observed in the inter-label distance distribution.

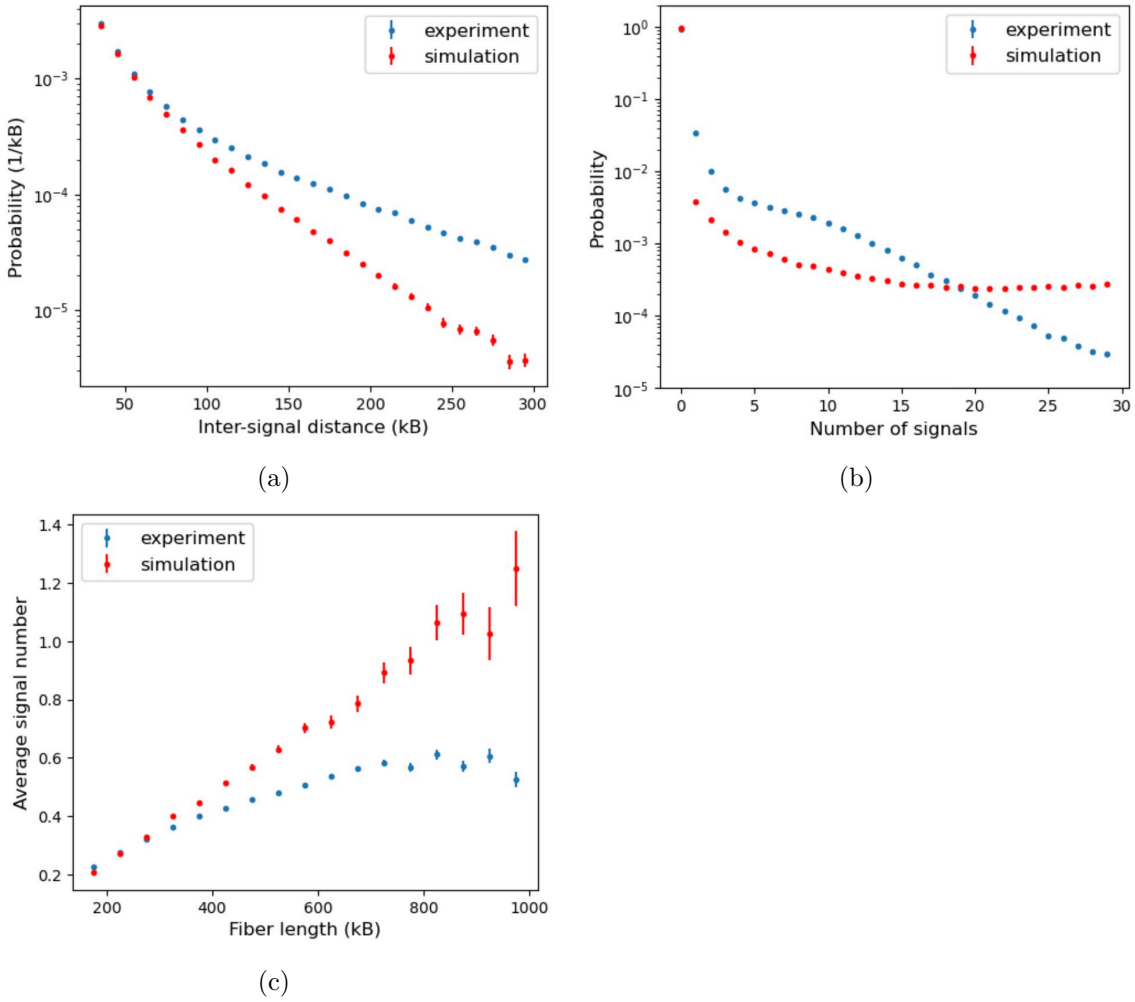


Figure 3.5: **Result of limited-fibre length model 2.** (Varying for I_0, r_0, and τ). (a) Inter-label distance distribution. (b) Distribution of number of labels. (c) Average label number as a function of fibre length.

3.3 Time-varying initiation model

A fundamental assumption in the previous models was that all initiation events occurred at $t = 0$. Here, we generalize our approach by considering initiations both at $t = 0$ due to aphidicolin blocks and at later time points due to an increasing initiation rate. In this context, we derive the model incorporating early initiations. We then compare the simulation results with the experimental data to test the updated model.

In deriving the equations for this model, we make the same assumptions as before. However, we now allow the total number of labels incorporating into the DNA to change over time because of the increasing number of active replication forks in the early replication process. Specifically, the incorporation rate becomes proportional to the number of labels inside the nucleus, $N(t)$. This adjustment allows us to account for the evolving dynamics of

replication events during the early stage of the cell cycle. The incorporation rate of Eq. 3.7 becomes

$$r(t) = k_i N(t), \tag{3.8}$$

where k_i is the incorporation constant. By taking $t = 0$ for the Eq. 3.8, we find

$$k_i = \frac{r_0}{N_0}, \tag{3.9}$$

where $N_0 \approx 1400$ is the total number of labels incubated inside the nucleus. This number can be estimated by counting the total number of incorporated labels per fold-coverage over the genome. We again assume that the decrease in the number of labels is a result only of label incorporation. Thus,

$$\dot{N}(t) = -2v n_f(t) r(t), \tag{3.10}$$

where $n_f(t)$ is the number of active replication forks. The fork velocity $2v$ determines the amount newly free nucleotides available for label incorporation (the factor 2 is because there are two DNA strands per replication fork). Using Eqs. 3.8 and 3.10, we find the incorporation rate

$$r(t) = -2v k_i n_f(t) r(t) \rightarrow r(t) = r_0 \exp\left\{-2v k_i \int_0^t \mathrm{d}t'\, n_f(t')\right\}. \tag{3.11}$$

According to Eq. 3.11, the incorporation rate is affected by the number of active replication forks from the start of the process. This result aligns with our intuition about the labeling realization, where the number of free labels decreases because of the active replication forks, which decreases the label-incorporation rate.

The number of active replication forks is determined from the number of initiations $n_+(t)$ and terminations $n_-(t)$ that occur after time $t = 0$. Thus,

$$n_f(t) = 2n_+(t) - 2n_-(t), \tag{3.12}$$

where the 2 comes from two forks coming from each initiation and two forks eliminating from each termination event. For early initiations, the number of terminations will be small

enough to ignore.[1] Therefore, number of early replication forks is

$$n_f(t) = 2\int_0^t \mathrm{d}t'\, I(t'). \tag{3.13}$$

Here, $I(t)\mathrm{d}t$ represents the average number of initiations occurring between t and $t+\mathrm{d}t$ all over the genome. Therefore, by understanding the initiation rate for early S phase, we can determine the incorporation rate. To incorporate this concept, we introduce a model for the initiation rate that combines aphidicolin-blocked initiation events ($t = 0$) and a linearly increasing initiation rate over time. This combined model accounts for both immediate initiations due to aphidicolin blocks (see Subsection 1.2.4) and the gradual rise in initiation events (see Subsection 1.3.1), allowing for a comprehensive understanding of the replication process.

$$I(t) = I_0\delta(t) + I_1 t. \tag{3.14}$$

By substituting Eqs. 3.14 and 3.13 into Eq. 3.11, we find

$$r(t) = r_0 \exp\left\{-2v\frac{r_0}{N_0}\left(I_0 t + \frac{I_1 t^3}{3}\right)\right\}. \tag{3.15}$$

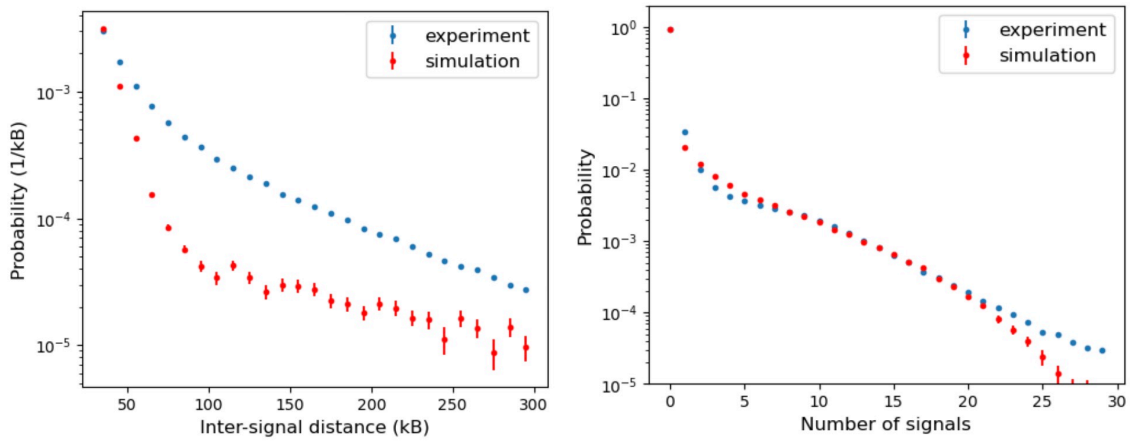
In Eq. 3.15, we adjusted the parameters I_0, I_1, and r_0 to match the probability densities. With the extra free parameter, the model's fit to the data improved, reducing the total loss function (as shown in Fig. 3.6). However, it continues to deviate significantly from the actual data. Specifically, the inter-label distance distribution displays a noticeable deviation from the observed pattern, unlike the realization with $t = 0$ initiations.

We also briefly investigated fitting using a loss function $\mathcal{L}$ calculated from log of the data

$$\mathcal{L} = \sum_i (\ln e_i - \ln o_i)^2, \tag{3.16}$$

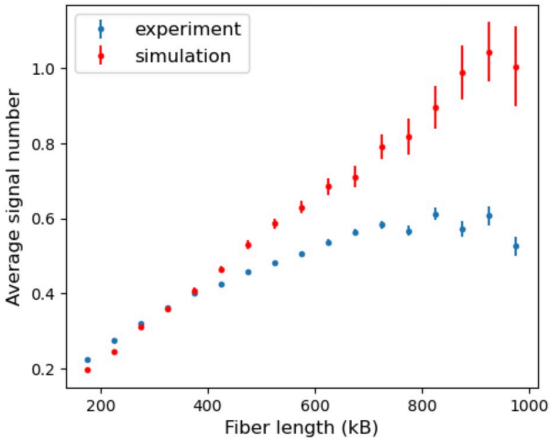
where e_i is the expected and o_i is the observed value of data point i. This loss function was motivated by the observation that the data cover a large range of magnitudes (see Fig. 3.1). However, the fit results were not qualitatively different from the χ^2 fits.

[1] The total number of initiations in human genome is approximately 50000 [11]. Thus, the average distance between each two initiation events would be $\frac{6.4\times10^6 \text{ kb}}{50000} = 128$ kb. An estimate for the average distance that replication forks span during labeling can be taken from the initiation zone sizes which are approximately 40 kb [10], which is relatively shorter than the average distance between the initiation events, implying that terminations are unlikely to occur during label incorporation. Please note that the number of early initiations is less than 50000.



(a)

(b)



(c)

Figure 3.6: **Results of time varying initiation rate model.** (a) Inter-label distance distribution. (b) Distribution of number of labels. (c) Average label number as a function of fibre length.

Chapter 4

Conclusion

In this thesis, we demonstrated that the exponential-decay model of [10], which described label-incorporation rates in ORM experiments and fits the inter-label distance distribution, falls short in explaining the distribution of the number of labels integrated into DNA fibers. This failure shows that the model developed in [10] cannot be used for inference of replication kinetics. In an attempt to improve the agreement with ORM data, we introduced a series of mathematical models tailored to ORM data, accounting for the shortcomings of the previous approach. The improvements included limiting the fibre length and accounting for the number of active replication forks. To implement these models, we simulated early S phase DNA replication kinetics. While the new models predict the expected average number of labels, there needs to be more work done to describe ORM data. We found that by using a linearly increasing initiation rate, we can explain the sparse nature of ORM data. However, the spacing between these labels does not agree with these models.

The previous analysis of [10] identifies initiation zones as regions with high label density; however, the main shortcoming of the ORM experiment is the sparse labeling, which complicates single-fibre inference and requires a more refined model for labeling. Several factors may contribute to the challenges encountered in this approach. DNA replication in human genomes is inherently stochastic, involving several enzymes and intricate chemical reactions at multiple locations along the DNA within the nucleus [5]. Several parameters, including DNA structure and genomic sequences, could potentially influence label incorporation. Additionally, using aphidicolin blocks interferes with the natural progression of S-phase by blocking some of the early initiation events. A challenge in processing the labels positions arises because multiple signals can potentially incorporate in a distance shorter than label position resolution (1 kb). Heterogeneity in initiation rates and variations in initiation rates between cells, may impact experimental data. Moreover, the incorporation of labels may exhibit asymmetry around initiation events owing to natural differences in nucleotide incorporation patterns between leading and lagging strands.

To enhance our models, one could consider localizing simulations to different genomic regions. Also, clustering might be a valuable strategy to distinguish individual initiation

events without the necessity to simulate replication kinetics. However, identifying multiple initiations that occur closely enough to influence cluster size poses a challenge. Moreover, employing more direct inference methods, such as machine-learning approaches could eliminate the need for accounting for the details of the experiment. An example can be training a neural network to predict initiation rate and then simulating the experimental data using the new initiation rate and recursively continuing this cycle until the initiation rate converges [87].

In summary, recent experiments have enhanced the labeling efficiency in ORM experiments by modifying the connection between nucleotide-like molecules and fluorescent dyes [10]. Using these datasets can enhance inference accuracy and serve as a robust method to test various inference approaches. The increased labeling efficiency provides sufficient data for making local inferences on initiation rates. Optical Replication Mapping represents an innovative technology, offering new insights into DNA replication kinetics in metazoan genomes. Finding a refined model for label incorporation rate would make it possible to infer the initiation position from single-fibre ORM data.

Bibliography

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. *Molecular Biology of the Cell*. Garland Science, Taylor and Francis Group, New York, NY, sixth edition, 2015.

[2] M. Goldberg, L. Hartwell, J. Fischer, and L. Hood. *Genetics: From Genes to Genomes*. McGraw-Hill Education, 2017.

[3] M. Hulke, D. Massey, and A. Koren. Genomic methods for measuring DNA replication dynamics. *Chromosome Research*, 28:49–67, March 2020.

[4] M. O'Donnell, L. Langston, and B. Stillman. Principles and Concepts of DNA Replication in Bacteria, Archaea, and Eukarya. *Cold Spring Harbor Perspectives in Biology*, 5(7):a010108–a010108, July 2013.

[5] J. Bechhoefer and N. Rhind. Replication timing and its emergence from stochastic processes. *Trends in Genetics*, 28(8):374–381, 2012.

[6] R. Berezney, Dharani D. Dubey, and Joel A. Huberman. Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. *Chromosoma*, 108(8):471–484, March 2000.

[7] M. Gauthier, P. Norio, and J. Bechhoefer. Modeling inhomogeneous DNA replication kinetics. *PLoS ONE*, 7(3):e32053, March 2012.

[8] S. Yang and J. Bechhoefer. How Xenopus laevis embryos replicate reliably: Investigating the random-completion problem. *Phys. Rev. E*, 78:041917, Oct 2008.

[9] O. Hyrien. Peaks cloaked in the mist: The landscape of mammalian replication origins. *Journal of Cell Biology*, 208(2):147–160, January 2015.

[10] W. Wang, K. Klein, K. Proesmans, H. Yang, C. Marchal, X. Zhu, T. Borrman, A. Hastie, Z. Weng, J. Bechhoefer, C. Chen, D. Gilbert, and N. Rhind. Genome-wide mapping of human DNA replication by optical replication mapping supports a stochastic model of eukaryotic replication. *Molecular Cell*, 81(14):2975–2988.e6, 2021.

[11] R. Milo and R. Phillips. *Cell Biology by the Numbers*. Garland Science, Taylor & Francis Group, New York, NY, 2016.

[12] J. Yang, J. Kim, S. Park, J. Jeon, Y. Shin, and S. Kim. Spatial and functional organization of mitochondrial protein network. *Scientific Reports*, 3(1):1403, March 2013.

[13] OpenStax College. DNA Nucleotides, April 2013. https://commons.wikimedia.org/wiki/File:DNA_Nucleotides.jpg, Website URL: https://commons.wikimedia.org, (accessed: 10.02.2024).

[14] S. Vengrova and J. Dalgaard, editors. *DNA Replication: Methods and Protocols*. Number 1300 in Methods in molecular biology. Humana Press, New York, second edition edition, 2015. OCLC: ocn903596278.

[15] A. Aze and D. Maiorano. Recent advances in understanding DNA replication: cell type–specific adaptation of the DNA replication program. *F1000Research*, 7:1351, August 2018.

[16] S. Ticau, L. Friedman, K. Champasa, I. Corrêa, J. Gelles, and S. Bell. Mechanism and timing of Mcm2-7 ring closure during DNA replication origin licensing. *Nature Structural & Molecular Biology*, 24(3):309–315, 2017.

[17] A. Costa and J. Diffley. The Initiation of Eukaryotic DNA Replication. *Annual Review of Biochemistry*, 91(1):107–131, June 2022.

[18] David O Morgan. Initiation of DNA replication, January 2007. https://commons.wikimedia.org/wiki/File:Initiation_of_DNA_replication.svg, Website URL: https://commons.wikimedia.org, (accessed: 10.02.2024).

[19] M. Fragkos, O. Ganier, P. Coulombe, and M. Méchali. DNA replication origin activation in space and time. *Nature Reviews Molecular Cell Biology*, 16(6):360–374, June 2015.

[20] P. Burgers and T. Kunkel. Eukaryotic DNA Replication Fork. *Annual Review of Biochemistry*, 86(1):417–438, June 2017.

[21] César Benito Jiménez. Elongation in DNA replication, April 2008. URL: https://commons.wikimedia.org/wiki/File:Replicación_bidireccional.jpg, Website URL: https://commons.wikimedia.org, (accessed: 10.02.2024).

[22] L. Balakrishnan and R. Bambara. Okazaki Fragment Metabolism. *Cold Spring Harbor Perspectives in Biology*, 5(2):a010173–a010173, February 2013.

[23] J. Dewar and J. Walter. Mechanisms of DNA replication termination. *Nature Reviews Molecular Cell Biology*, 18(8):507–516, August 2017.

[24] S. Chodavarapu and J. Kaguni. Chapter One - Replication Initiation in Bacteria. In Laurie S. Kaguni and Marcos Túlio Oliveira, editors, *DNA Replication Across Taxa*, volume 39 of *The Enzymes*, pages 1–30. Academic Press, 2016.

[25] C. Nieduszynski, S. Hiraga, P. Ak, C. Benham, and A. Donaldson. OriDB: a DNA replication origin database. *Nucleic Acids Research*, 35(suppl_1):D40–D46, January 2007.

[26] S. Yang, N. Rhind, and J. Bechhoefer. Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. *Molecular Systems Biology*, 6(1):404, January 2010.

[27] Shankar P. Das, Tyler Borrman, Victor W.T. Liu, Scott C.-H. Yang, John Bechhoefer, and Nicholas Rhind. Replication timing is regulated by the number of MCMs loaded at origins. *Genome Research*, 25(12):1886–1892, December 2015.

[28] J. Hamlin, L. Mesner, O. Lar, R. Torres, S. Chodaparambil, and L. Wang. A revisionist replicon model for higher eukaryotic genomes. *Journal of Cellular Biochemistry*, 105(2):321–329, October 2008.

[29] A. Demczuk, M. Gauthier, I. Veras, S. Kosiyatrakul, C. Schildkraut, M. Busslinger, J. Bechhoefer, and P. Norio. Regulation of DNA Replication within the Immunoglobulin Heavy-Chain Locus During B Cell Commitment. *PLoS Biology*, 10(7):e1001360, July 2012.

[30] M. Meselson and F. Stahl. The replication of DNA in *Escherichia coli*. *Proceedings of the National Academy of Sciences*, 44(7):671–682, July 1958.

[31] R Hand. Deoxyribonucleic acid fiber autoradiography as a technique for studying the replication of the mammalian chromosome. *Journal of Histochemistry & Cytochemistry*, 23(7):475–481, July 1975.

[32] G. Pagoulatos and M. Yaniv. High resolution two-dimensional electrophoresis of proteins bound to heterogeneous nuclear RNA. *FEBS Letters*, 74(1):115–120, February 1977.

[33] P. Zhao, T. Sasaki, and D. Gilbert. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biology*, 21(1):76, December 2020.

[34] T. Ryba, D. Battaglia, B. Pope, I. Hiratani, and D. Gilbert. Genome-scale analysis of replication timing: from bench to bioinformatics. *Nature Protocols*, 6(6):870–895, June 2011.

[35] J. Rivera-Mulia, C. Trevilla-Garcia, and S. Martinez-Cifuentes. Optimized Repli-seq: improved DNA replication timing analysis by next-generation sequencing. *Chromosome Research*, 30(4):401–414, December 2022.

[36] V. Dileep and D. Gilbert. Single-cell replication profiling to measure stochastic variation in mammalian replication timing. *Nature Communications*, 9(1):427, January 2018.

[37] X. Liao, M. Makris, and X. Luo. Fluorescence-activated Cell Sorting for Purification of Plasmacytoid Dendritic Cells from the Mouse Bone Marrow. *Journal of Visualized Experiments*, 54641(117):54641, November 2016.

[38] A. Koren, R. Handsaker, N. Kamitaki, R. Karlić, S. Ghosh, P. Polak, K. Eggan, and S. McCarroll. Genetic Variation in Human DNA Replication Timing. *Cell*, 159(5):1015–1026, November 2014.

[39] I. Hiratani and S. Takahashi. DNA Replication Timing Enters the Single-Cell Era. *Genes*, 10(3):221, March 2019.

[40] S. Gnan, J. Josephides, X. Wu, M. Spagnuolo, D. Saulebekova, M. Bohec, M. Dumont, L. Baudrin, D. Fachinetti, S. Baulande, and C. Chen. Kronos scRT: a uniform framework for single-cell replication timing analysis. *Nature Communications*, 13(1):2329, April 2022.

[41] D. Massey and A. Koren. High-throughput analysis of single human cells reveals the complex nature of DNA replication timing control. *Nature Communications*, 13(1):2402, May 2022.

[42] B. Pope and D. Gilbert. The Replication Domain Model: Regulating Replicon Firing in the Context of Large-Scale Chromosome Architecture. *Journal of Molecular Biology*, 425(23):4690–4695, November 2013.

[43] B. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. Vera, Y. Wang, R. Hansen, T. Canfield, R. Thurman, Y. Cheng, G. Gülsoy, J. Dennis, M. Snyder, J. Stamatoyannopoulos, J. Taylor, R. Hardison, T. Kahveci, B. Ren, and D. Gilbert. Topologically associating domains are stable units of replication-timing regulation. *Nature*, 515(7527):402–405, November 2014.

[44] A. Bracci, A Dallmann, Q. Ding, M. Hubisz, M. Caballero, and A. Koren. The evolution of the human DNA replication timing program. *Proceedings of the National Academy of Sciences*, 120(10):e2213896120, 2023.

[45] Robert Schleif. DNA LOOPING. *Annual Review of Biochemistry*, 61(1):199–223, June 1992.

[46] H. Técher, S. Koundrioukoff, D. Azar, T. Wilhelm, S. Carignon, O. Brison, M. Debatisse, and B. Le Tallec. Replication Dynamics: Biases and Robustness of DNA Fiber Analysis. *Journal of Molecular Biology*, 425(23):4845–4855, November 2013.

[47] P. Norio, S. Kosiyatrakul, Q. Yang, Z. Guan, N. Brown, S. Thomas, R. Riblet, and C. Schildkraut. Progressive Activation of DNA Replication Initiation in Large Domains of the Immunoglobulin Heavy Chain Locus during B Cell Development. *Molecular Cell*, 20(4):575–587, November 2005.

[48] C. Müller, M. Boemo, P. Spingardi, B. Kessler, S. Kriaucionis, J. Simpson, and C. Nieduszynski. Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. *Nature Methods*, 16(5):429–436, May 2019.

[49] N. Tanner, J. Loparo, S. Hamdan, S. Jergic, N. Dixon, and Antoine M. Van O. Real-time single-molecule observation of rolling-circle DNA replication. *Nucleic Acids Research*, 37(4):e27–e27, March 2009.

[50] X. Michalet, R. Ekong, F. Fougerousse, S. Rousseaux, C. Schurra, N. Hornigold, M. Slegtenhorst, J. Wolfe, S. Povey, J. Beckmann, and A. Bensimon. Dynamic Molecular Combing: Stretching the Whole Human Genome for High-Resolution Studies. *Science*, 277(5331):1518–1523, September 1997.

[51] E. Schwob, C. de Renty, V. Coulon, T. Gostan, C. Boyer, L. Camet-Gabut, and C. Amato. Use of DNA Combing for Studying DNA Replication In Vivo in Yeast and Mammalian Cells. *Methods in molecular biology (Clifton, N.J.)*, 521:673–87, 02 2009.

[52] G. Guilbaud, A. Rappailles, A. Baker, C. Chen, A. Arneodo, A. Goldar, Y. d'Aubenton Carafa, C. Thermes, B. Audit, and O. Hyrien. Evidence for Sequential and Increasing Activation of Replication Origins along Replication Timing Gradients in the Human Genome. *PLoS Computational Biology*, 7(12):e1002322, December 2011.

[53] W. John Kress and David L. Erickson, editors. *DNA Barcodes: Methods and Protocols*. Number 858 in Methods in Molecular Biology. Humana Press/Springer, New York, 2012.

[54] F. De Carli, N. Menezes, W. Berrabah, V. Barbe, A. Genovesio, and O. Hyrien. High-Throughput Optical Mapping of Replicating DNA. *Small Methods*, 2(9):1800146, 2018.

[55] E. Lam, A. Hastie, C. Lin, D. Ehrlich, S. Das, M. Austin, P. Deshpande, H. Cao, N. Nagarajan, M. Xiao, and P. Kwok. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature Biotechnology*, 30(8):771–776, August 2012.

[56] T. Murate, T. Hotta, K. Tsushita, M. Suzuki, T. Yoshida, S. Saga, H. Saito, and S. Yoshida. Aphidicolin, an inhibitor of DNA replication, blocks the TPA-induced differentiation of a human megakaryoblastic cell line, MEG-O1. *Blood*, 78(12):3168–3177, December 1991.

[57] J. Gehl. Electroporation: theory and methods, perspectives for drug delivery, gene therapy and research. *Acta Physiologica Scandinavica*, 177(4):437–447, April 2003.

[58] E. Knierim, B. Lucke, J. Schwarz, M. Schuelke, and D. Seelow. Systematic Comparison of Three Methods for Fragmentation of Long-Range PCR Products for Next Generation Sequencing. *PLoS ONE*, 6(11):e28240, November 2011.

[59] Y. Zhang, X. Kong, L. Gao, Y. Tian, L. Wen, and L. Jiang. Fabrication of Nanochannels. *Materials*, 8(9):6277–6308, 2015.

[60] A. De Moura, R. Retkute, M. Hawkins, and C. Nieduszynski. Mathematical modelling of whole chromosome replication. *Nucleic Acids Research*, 38(17):5623–5633, September 2010.

[61] A. Kaykov and P. Nurse. The spatial and temporal organization of origin firing during the S-phase of fission yeast. *Genome Research*, 25(3):391–401, March 2015.

[62] O. Hyrien and A. Goldar. Mathematical modelling of eukaryotic DNA replication. *Chromosome Research*, 18(1):147–161, January 2010.

[63] S. Yang and J. Bechhoefer. How *Xenopus laevis* embryos replicate reliably: Investigating the random-completion problem. *Physical Review E*, 78(4):041917, October 2008.

[64] A. Kolmogorov. On the Statistical Theory of Metal Crystallization. *Izv. Akad. Nauk SSSR, Ser. Math*, 1:335–360, 1937.

[65] W. Johnson and R. Mehl. Reaction kinetics in processes of nucleation and growth. *Trans. Am. Inst. Min. Metall. Eng.*, 135:416–442, 1939.

[66] M. Avrami. Granulation, Phase Change, and Microstructure Kinetics of Phase Change. III. *The Journal of Chemical Physics*, 9(2):177–184, 1941.

[67] K. Sekimoto. Kinetics of magnetization switching in a 1-D system II-long time behavior of switched domains. *Physica A: Statistical Mechanics and its Applications*, 128(1):132–149, 1984.

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

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

[70] S. Jun, J. Herrick, A. Bensimon, and J. Bechhoefer. Persistence Length of Chromatin Determines Origin Spacing in Xenopus Early-Embryo DNA Replication: Quantitative Comparisons between Theory and Experiment. *Cell Cycle*, 3(2):211–217, February 2004.

[71] E. Mirkin and S. Mirkin. Replication Fork Stalling at Natural Impediments. *Microbiology and Molecular Biology Reviews*, 71(1):13–35, March 2007.

[72] R. Yousefi and M. Rowicka. Stochasticity of replication forks' speeds plays a key role in the dynamics of DNA replication. *PLOS Computational Biology*, 15(12):e1007519, December 2019.

[73] C. Conti, B. Saccà, J. Herrick, C. Lalou, Y. Pommier, and A. Bensimon. Replication Fork Velocities at Adjacent Replication Origins Are Coordinately Modified during DNA Replication in Human Cells. *Molecular Biology of the Cell*, 18(8):3059–3067, August 2007.

[74] Y. Gindin, M. Valenzuela, M. Aladjem, P. Meltzer, and S. Bilke. A chromatin structure-based model accurately predicts replication timing in human cells. *Molecular Systems Biology*, 10(3):722, March 2014.

[75] J. Herrick, S. Jun, J. Bechhoefer, and A. Bensimon. Kinetic Model of DNA Replication in Eukaryotic Organisms. *Journal of Molecular Biology*, 320(4):741–750, July 2002.

[76] Y. Gindin, P. Meltzer, and S. Bilke. Replicon: a software to accurately predict DNA replication timing in metazoan cells. *Frontiers in Genetics*, 5, November 2014.

[77] S. Jun and N. Rhind. Just-in-time DNA replication. *Physics*, 1:32, October 2008.

[78] J. Arbona, A. Goldar, O. Hyrien, A. Arneodo, and B. Audit. The eukaryotic bell-shaped temporal rate of DNA replication origin firing emanates from a balance between origin activation and passivation. *eLife*, 7:e35192, jun 2018.

[79] A. Goldar, H. Labit, K. Marheineke, and O. Hyrien. A Dynamic Stochastic Model for DNA Replication Initiation in Early Embryos. *PLoS ONE*, 3(8):e2919, August 2008.

[80] M. Gauthier and J. Bechhoefer. Control of DNA Replication by Anomalous Reaction-Diffusion Kinetics. *Physical Review Letters*, 102(15):158104, April 2009.

[81] D. Bhat, S. Hauf, C. Plessy, Y. Yokobayashi, and S. Pigolotti. Speed variations of bacterial replisomes. *eLife*, 11:e75884, July 2022.

[82] M. Gauthier, J. Herrick, and J. Bechhoefer. Defects and DNA Replication. *Physical Review Letters*, 104(21):218104, May 2010.

[83] H. Zhang and J. Bechhoefer. Reconstructing DNA replication kinetics from small DNA fragments. *Physical Review E*, 73(5):051903, May 2006.

[84] A. Baker, B. Audit, S. Yang, J. Bechhoefer, and A. Arneodo. Inferring Where and When Replication Initiates from Genome-Wide Replication Timing Data. *Physical Review Letters*, 108(26):268101, June 2012.

[85] A. Baker and J. Bechhoefer. Inferring the spatiotemporal DNA replication program from noisy data. *Physical Review E*, 89(3):032703, March 2014.

[86] A. Bazarova, C. Nieduszynski, I. Akerman, and N. Burroughs. Bayesian inference of origin firing time distributions, origin interference and licencing probabilities from Next Generation Sequencing data. *Nucleic Acids Research*, 47(5):2229–2243, 02 2019.

[87] J. Arbona, H. Kabalane, J. Barbier, A. Goldar, O. Hyrien, and B. Audit. Neural network and kinetic modelling of human genome replication reveal replication origin locations and strengths. *PLOS Computational Biology*, 19(5):e1011138, May 2023.

[88] E. Ben-Naim and P. Krapivsky. Nucleation and growth in one dimension. *Phys. Rev. E*, 54:3562–3568, Oct 1996.

[89] M. Pincus. Letter to the Editor—A Monte Carlo Method for the Approximate Solution of Certain Types of Constrained Optimization Problems. *Operations Research*, 18(6):1225–1228, 1970.

[90] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[91] Y Xiang, D.Y Sun, W Fan, and X.G Gong. Generalized simulated annealing algorithm and its application to the thomson model. *Physics Letters A*, 233(3):216–220, 1997.

[92] M. Hardt and B. Recht. *Patterns, Predictions, and Actions: Foundations of Machine Learning*. Princeton University Press, Princeton, 2022.

[93] M. Fu. Feature article: Optimization for simulation: Theory vs. practice. *INFORMS Journal on Computing*, 14(3):192–215, 2002.