Genome-Wide Mapping of Human DNA Replication by Optical Replication Mapping Supports a Stochastic Model of Eukaryotic Replication

Weitao Wang

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée. Institut Curie.
Dans le cadre d'une cotutelle avec Sorbonne Université Faculté des Sciences École Doctorale ED515 Complexité du Vivant (CdV).

La cartographie génomique de la réplication de l'ADN de cellules humaines par la cartographie de réplication optique soutiennent un modèle stochastique de réplication chez eucaryote

Genome-Wide Mapping of Human DNA Replication by Optical Replication Mapping Supports a Stochastic Model of Eukaryotic Replication

Soutenue par
Weitao Wang
Le 15 Juillet 2021

École doctorale n° 515
Complexité du Vivant

Spécialité
Génomique, bioinformatique

Composition du jury :

Sarah Lambert Directrice de recherche, PSL-I. Curie, Orsay	Présidente
Benjamin Audit Directeur de recherche, ENS de Lyon	Rapporteur
Yea-Lih Lin Directrice de recherche, IGH, Montpelier	Rapportrice
Chun-Long Chen Chargé de recherche, PSL-I. Curie, Paris	Directeur de thèse
Nick Rhind Professor, Umass, USA	Examinateur
Morgane Thomas-Chollier Maître de conférences, PSL-ENS Paris	Examinatrice
Torsten Krude Professor, Cambridge, UK	Examinateur

institutCurie | PSL

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

Remerciement / Acknowledgements

Time flies like an arrow, three and a half years is fleeting, and it is time for graduation. During the entire Ph.D. stage, I am most grateful to my mentor, Dr Chun-long Chen. His profound knowledge, rigorous academic spirit, and tireless and noble teacher ethics have had a profound impact on me. It not only allowed me to establish ambitious academic goals, but also made me understand a lot of truths about how to behave in the world. At the same time, it also gave me meticulous care in my spirit and life. Many times, Dr Chunlong Chen helped me revise the paper until late at night. When I woke up the next day, he was still working and patiently explained the problems in my paper. When the progress of my project was blocked, he always provided clever advice and guidance at the first time, allowing me to overcome each difficulty. In addition, the successful completion of this thesis is inseparable from the care and help of the cooperating teachers. I would also like to thank Professor Nick, Professor David Gilbert, and Professor John Bechhoefer for their guidance and help. They have done a lot of work on this topic. Thank you very much! In my life and study, I have also received care and help from all the students and members in the laboratory. Thanks, Stefano, Manuela, Joseph and Yaqun, you have spent a very happy time with me! Finally, I would also like to thank my mother who cultivated me and support me all the time! I really appreciate all of your selfless help and care from the bottom of my heart.

Table of Contents

List of abbreviations

2D/3D: 2 Dimensions / 3 Dimensions

ARSs: Autonomously Replicating Sequences

AUC: area under the ROC curve

BrdU: 5-bromo2-deoxyuridine

CDC45: Cell Division Cycle 45

CDC6: Cell Division Cycle 6

CDT1: CDC10-dependent transcript 1

CGIs: CpG islands

ChIP: Chromatin immunoprecipitation

CHK1: checkpoint kinase 1

DLE-1: direct labeling enzyme 1

DLS: Direct Label and Stain

DNase I: DNase I hypersensitive sites

EdU: 5-ethynyl-20-deoxyuridine

ENCODE: The Encyclopedia of DNA Elements

FACS: fluorescence-activated cell sorter

FDI_RFD: replication fork directionality calculated by ORM segments with +/- FDI

FDI: fork direction index

FISH: Fluorescent in situ hybridization

FPR: false positive rate

G4: G-quadruplex

GINS: go-ichi-ni-san

GMM: Gaussian mixture model

HOMARD: high-throughput optical mapping of replicating DNA

HU: hydroxyurea

IZ: initiation zone

LOESS: locally estimated scatterplot smoothing

MCM: mini-chromosome maintenance

NLRS: Nick Label Repair and Stain

NGS: next generation sequencing

Nt.BspQI: BspQI is a thermostable Type IIS restriction endonuclease and Nt.BspQI is nicking endonuclease

ORC: origin recognition complex

ORIs: origin replication initiation sites

ORM: Optical replication mapping

PBS: Phosphate-buffered saline

PCNA: proliferating cell nuclear antigen

Pol ε: DNA polymerase ε

Pre-IC: pre-initiation complex

Pre-RC: pre-replication complex

RECQL4: ATP-dependent DNA helicase Q4

RFC: replication factor C

RFD: replication fork directionality

RGP: red ORM signal genomic position

ROC curve: receiver operating characteristic curve

RPA: replication protein A

SNR: signal-noise ratio

SNS: short nascent strand

SSS: short single-strands

SVM: supported vector machine

TAD: Topologically associating domains

TOPBP1: DNA topoisomerase 2-binding protein 1

TPR: true positive rate

TSSs: transcription start sites

Xgboost: eXtreme Gradient Boosting

CHAPTER 1

Introduction

DNA replication is the basis of biological inheritance in all living beings. The process gives rise to two identical copies of the original DNA molecule. Its importance is self-evident, and the deregulation of replication can challenge genome stability and lead to mutations, cancer, and many other genetic diseases. Because of its importance, the study of replication mechanisms has been widely concerned by biologists. DNA replication begins at specific locations in the genome, and the unwinding of DNA occurs with the help of the enzyme helicase and gyrase in almost all DNA replication processes. However, the mechanism of replication initiation varied across the species. In bacteria, the DnaA protein determines the site of initiation on the genome, and replication origins share highly specific homologous sequences (Masai et al., 2010). However, in the eukaryotic cell, the initiation process is triggered by the origin recognition complex (ORC). For budding yeast, the ORC binding with specific origin sites, which contain 17 bp conserved sequences, such as TTTTTTATGTTTGTT (Eaton et al., 2010; Nieduszynski, 2016; Theis et al., 1999). But for most eukaryotes, there is no ORC consensus site with a known motif being identified. There are several clues being identified:

a. In fission yeast and *Drosophila*, origin selection prefers AT-rich intergenetic regions. (Chuang and Kelly, 1999; Schaarschmidt et al., 2004; Vashee et al., 2003)

b. Origin recognition is related to DNA topology: topoisomerases have been associated with human replication origin (Abdurashidova et al., 2007) and supercoiled DNA tends to become origin in fission yeast and *Drosophila* (Houchens et al., 2008; Remus et al., 2004).

c. The transcription factors with Myb protein may facilitate the ORC site-specific feature at ACE3 and Ori-β (Beall et al., 2002).

Even so, these conclusions are specific to specific loci or species and are not absolutely valid in other species. For example, the AT-rich feature is not obvious in mammalian ORC *in vitro* (Vashee et al., 2003; De Carli et al., 2018). Such interspecies difference also increases the difficulty of locating the replication origins, especially for origin detection in human cells.

Why is studying replication initiation important for exploring the replication mechanism? The two most important basic concepts in the DNA replication process are replication initiation time and location, in other words, when a given genomic region being replicated and where the DNA replication process starts. They are the two direct elements that determine the replication process. Since the origin sequence of mammals is almost irregular, in recent years, a large number of studies have been directed at ORC (Sugimoto et al., 2018), or on the sequence generated with the initiation of replication, such as short nascent DNA, Okazaki fragment (Petryk et al., 2016; Picard et al., 2014) and tried to infer the location of origin by derivative of initiation. However, despite intensive studies, the mechanisms that coordinate where and when replication initiates in the human genome remain poorly known. This is due to the low efficiency of initiation, which only shows up in 1%~10%

cells (Demczuk et al., 2012; Dijkwel et al., 2002), and heterogeneous selection of replication origins from cell to cell, which makes the data obtained by classical population-based approaches is very controversial (Langley et al., 2016; Mesner et al., 2013)

A potential way to solve the problem is a single-molecular method with strong monitoring sensitivity. At the time I started my Ph.D. thesis, the only known two single-molecule methods are DNA combing and SMARD, both with a few hundred fibers throughput, cannot meet the requirements of genome-wide sequencing, no matter in coverage or depth.

In this context, the present work aims to develop new cutting-edge high-throughput genomic approaches to study the spatio-temporal replication program of the human genome at the single-molecule level. Together with the genome-wide data analyses, I aim to address the following questions: What determines the replication program, i.e. the position, the time of firing, and the efficiency of replication origins, in the human genome?

1.1 DNA replication mechanism and the corresponding knowledge

1.1.1 Cell Cycle

Life is a continuous process passed from one generation to the next, so it is a process of continuous renewal and continuous starting from the beginning. A cell was born from the division of its parent cell and ends with the formation of its daughter cells or the cell's own death. The entire cell cycle is divided into two stages: interphase (including G1, S, G2 phase) and split period (M phase) (Fig. 1.1). During this process, the genetic material of the cell is replicated within the S phase and distributed equally to the two daughter cells in the M phase.

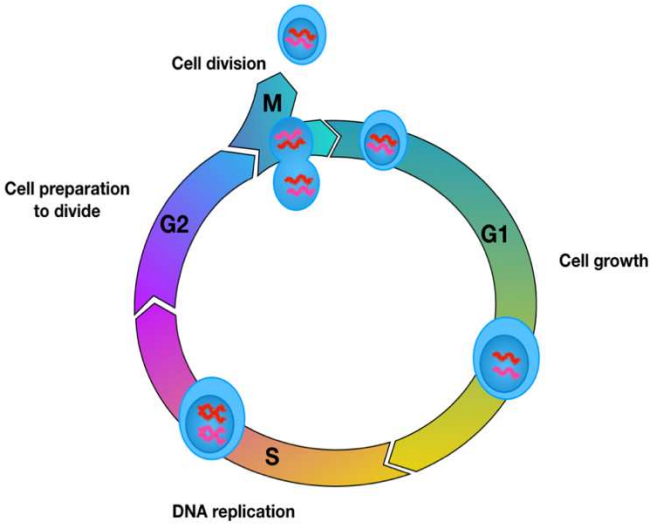


Figure 1.1: Schematic of the cell cycle. The entire cell cycle is divided into two stages: interphase (including G1, S, G2 phase) and split period (M phase). Cells begin to grow in the G1 phase, then enter into the S phase for DNA synthesis. Followed by the G2 phase when continuing for cell growth and preparations for mitosis. After that, the final step of the cell circle is cell division in the M phase. The length of the arrow represents the time ratio in a cell circle for a standard human somatic cell. But the ratio is various in different cell types, the ES cells have a shorter G1 phase, and cancer cells can also present some alteration in the cell cycle control.

1.1.2 Replication Origins

DNA replication starts at specific positions, we call these positions replication origins. As mentioned above, the characteristics of origins have strong interspecies differences. For example, in budding yeast and bacteria genome, the replication origins are located at conserved motif sequences. But for most mammals, the sequence of origin is totally irregular, even if the fission yeast, which belongs to the homologous unicellular organism to budding yeast (Toya et al., 1999) also just prefers to fire origins in AT-rich regions. At present, the characteristics of mammalian replication initiation sites recognized by mainstream academic circles are only the following facts.

a. Mammalian ORC DNA binding has non-sequence-specific nature.
b. Initiation is enriched in euchromatic promoter and enhancer regions, consistent with its correlation with accessible chromatin (Cayrou et al., 2015; Ganier et al., 2019; Petryk et al., 2016; Pourkarimi et al., 2016)
c. Initiation in heterochromatin appears to be even more heterogeneous, making heterochromatic initiation sites even less well understood (Cayrou et al., 2015; Petryk et al., 2016)
d. Some origins only occurred when cells are confronted with replication stress, for example, the frequent fork stalling in cancer cells makes some origins, which are silent or late replicated in normal cell lines, fire early (Macheret and Halazonetis, 2018)

The replication initiation sites identified in the current study from Hela cells meet all the above characteristics (see below sections for detail). The sites of human replication initiation are not confined to well-defined replication origins but are instead the origins distributed across specific initiation zones. Also, they are highly enriched in promoter and enhancer regions, and we found some early replicated origins within late replicating regions defined by population-based replication timing values. In subsequent chapters, I will elaborate on these in more detail.

1.1.3 Replication unit

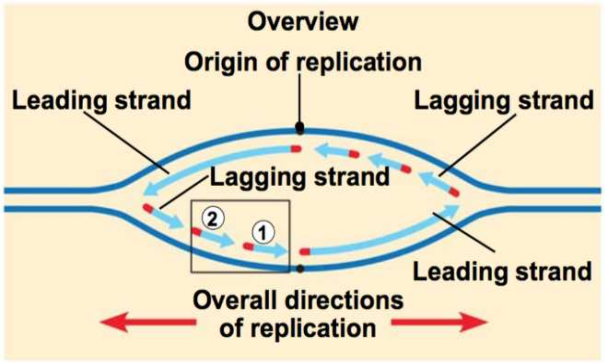


Figure 1.2: Schematic diagram of a replication bubble. Since the replication direction could only be from 5' to 3' end like the blue arrows showing from red ends to blue, the two strands according to the principle of complementary have opposite replication direction. The continuous synthetic strand follows the overall direction of the replication fork (red arrow) with the same orientation called leading strand and the other named lagging strand with inverse orientation, has to replicate intermittently by multi short Okazaki fragments like the ①, ② in the square. (The picture is from an online biological video course: CLUCH).

Once an origin started to fire, the unwinding of DNA double strands will produce a bubble-like structure called a replication bubble (Fig. 1.2). This is the smallest unit of the DNA replication process. The replication bubble grows in two directions from its origin. There are two Y shape structures named replication forks in a replication bubble, which is where the parental DNA double helix is split. Normally, the replication direction means the forward direction of a replication fork. And along with the movement of the replication forks, using the two untwisted DNA strands as a template, new sub-strands are synthesized from 5' to 3' respectively.

1.1.4 The complete biological replication initiation process

This section explains in detail the entire process from the formation of all components to origin activation. All steps are roughly divided into 3 processes: Origin licensing, Pre-IC (pre-initiation complex) formation, and origin firing (Fig. 1.3). Firstly, in the G1 phase, the ORC binding to the origin and recruit CDC6 (Cell Division Cycle 6), CDT1 (CDC10-dependent transcript 1), and MCM (mini-chromosome maintenance) to form Pre-RC (pre-replication complex) (Fig. 1.3a). This process is called origin licensing. It is worth mentioning that MCM named helicase complex, which contains the six subunits MCM2–7. It is the last step of the licensing reaction and can take place only if ORC, CDC6, and CDT1 are already bound to origins. In some methods of detecting origin, MCM is often used as the target protein. But the licensing origins may not go through the following processes afterward, they are just potential origins waiting for the catalysis of CDC6 and DDK to promote the following process.

Then during the period G1 to S, there are still a lot of small-molecule proteins converge on part of licensing origins. Pre-IC is further generated on the basis of pre-RC successively (Fig. 1.3b). This step is triggered by DBF4-dependent kinase (DDK) and cyclin-dependent kinases (CDKs) at the G1–S phase transition. They are not the component of Pre-IC, but DDK and CDKs phosphorylate several replication factors including MCM10, CDC45 (Cell Division Cycle 45), RECQL4 (ATP-dependent DNA helicase Q4), treslin, GINS (an acronym created from the first letters of the Japanese numbers 5-1-2-3, i.e. go-ichi-ni-san, in a reference to the 4 protein subunits of the complex: Sld5, Psf1, Psf2, and Psf3), TOPBP1 (DNA topoisomerase 2-binding protein 1), and Pol ε (DNA polymerase ε) to promote their loading on origins. In addition, DDK and CDKs phosphorylate the residues of the MCM2–7 complex, which leads to helicase activation and DNA unwinding.

Once helicase activation is triggered, marking the beginning of origin firing (Fig. 1.3c). The helicase activation makes MCM2–7 double hexamer divides into two hexamers that acting on the two replication forks emit from the same replication origin. Meanwhile, helicase activation induces the recruitment of other proteins, such as RFC (replication factor C), PCNA (proliferating cell nuclear antigen), RPA (replication protein A), and other DNA polymerases that convert the pre-IC into two functional replisomes at two opposite moving replication forks. The functional helicase at the forks is the CMG complex inside the dashed line square (which contains CDC45, the MCM hexamer, and the GINS complex). In population-based data, the selection of origins can be very different from cell to cell; thus, these flexible origins make the replication mechanism more elusive. Inhibition of adjacent origins within a replication unit is controlled in part by the checkpoint kinases Ser/Thr protein kinase ATR and Ser protein kinase ATM that activate checkpoint kinase 1 (CHK1) and CHK2 (Fragkos et al., 2015a). However, the exact mechanisms that are responsible for the local inhibition of these flexible origins remain unclear. Similarly, it has not yet been determined how flexible origins are selected for activation or silencing.

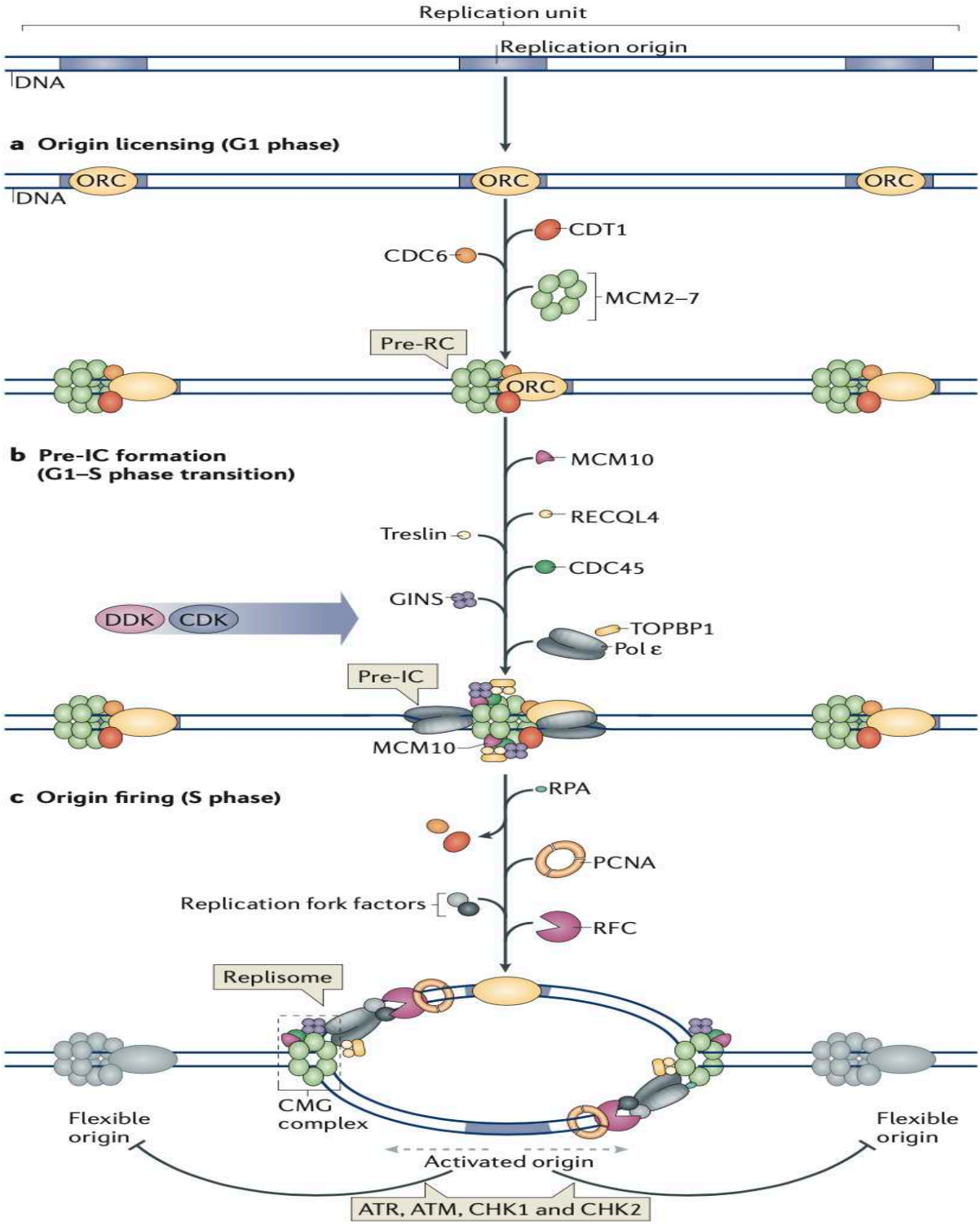


Figure 1.3: Complete replication initiation process model figure from (Fragkos et al., 2015b). Please see the main text (section 1.1.3) for a detailed explanation.

1.1.5 Replication timing

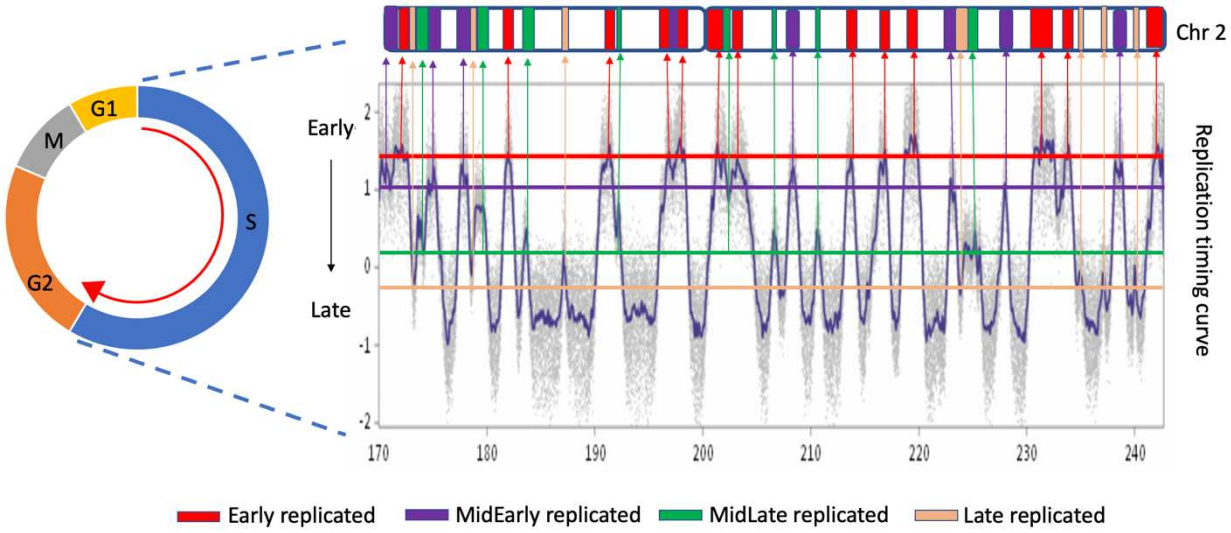


Figure 1.4: Schematic of replication timing: The left circle represents the cell circle and the dashed line marking the S phase to show the right panel on replicated process occurs in the S phase. On the right, the bottom plot is a replication timing curve along a genomic region. The X-axis is the genomic position in megabase (Mb) along with chromosome 2 corresponding to the chromosome pictogram on the top. The Y-axis prensents the replication timing obtained by log2(E/L) ratio. The higher the timing value being on the Y-axis, the corresponding position will be replicated earlier. The raw bulk data of replication timing could be very noisy with big variance showed by discrete gray dots. The timing curve is a fitting profile, which reflects the average timing value in most cells corresponding to a specific genomic position. It is a comprehensive result after referring to population-based data. For example, for the peak site with the highest timing value in the figure, it can only be explained that according to statistics, this site is selected as the firing origin earlier in most cells and replicated at the beginning of the S phase, but it does not rule out that a small number of cells are in the early stage, while the origin site is not activated after origin licensing, resulting in the site replicated late in these cells. Here we set 4 horizontal cutoff lines with red, purple, green, and fleshed color, which represent the temporal order from early to late in the turn. At the moment corresponding to the 4 cutoff lines, I have marked the corresponding replicated chromosome region as the corresponding color and selected some regions to show the corresponding relationship between the timing curve and the region with arrows of corresponding colors. In each period of the S phase, each cutoff horizontal line always intersects with a part of replication curves, which means the position corresponding to the abscissa of the intersection is being replicated at this moment to the cutoff timing value, and the curve above the cutoff line is the area that has been replicated, and the part below it is the area that has not been replicated yet. With the passage of time in the S phase, which can be imagined as a gradually decreasing horizontal line. The line's timing value drops from 2 to -2 and passes through the moments corresponding to the 4 cutoff values that I have marked. For a certain local peak in the timing curve, the horizontal line must be the first to touch the top of the peak during the descending process. These peaks represent the corresponding X positions of the chromosome, in most cells, is the activated firing origin at the moment corresponding to the peaks' Y timing values. The opposite replication forks generated here cause the positions on the left and right sides of the site to be replicated sequentially, so a peak shape is formed in the timing curve. And there are firing origins that show up successively in a different period from early to late during the S phase.

In eukaryotic cells, replication timing is a value to describe the temporal order when ongoing DNA replication along the chromosome arrives in one genomic position. Along the entire genome, some positions start to replicate at the beginning of the S phase, like the origins in the early S phase. With time passing, more and more origins started to fire, and when facing replication forks from two adjacent replication origins meet each other, the replication termination occurs. This

replication process will keep on going until the whole genome finishing replicating. And all bases on the genome should be used as a value to represent the timing when the replication fork arrives in this position. So as to achieve one-to-one correspondence in timing and position. This value is specified in one given cell, but normally varies from one cell to another because of a different selection of licensing origins. Different methods can be used to define the replication timing value at the population level or at the single-cell level (Dileep and Gilbert, 2018), and the specific replication profile is based on the cell line and cell growth condition. Different cell types and replication stress conditions will result in different replication origin numbers and selection causing corresponding replication timing curves to change. But for cell population in a given cell type and condition, replication timing values of a specific position could be various from cell to cell, however, for a given position, majority of cells will still replicate within a close range window (~2 h around the average replication timing) within S phase (Dileep and Gilbert, 2018). Based on different methods used, the replication timing values can be either defined from -1 to 1 by log 2 ratio of the amount of newly replicated DNA detected in early and late S phases at different genomic locations along the genome (Gilbert, 2010) or classify the entire S phase into 4-6 different periods from early to late and calculate the 50% of a given genomic region has been replicated to get the S50 values (Chen et al., 2010; Dellino et al., 2013). I will provide below, a detailed description of both methods.

1.1.5.1 Determination of replication timing by log2 ratio of replicated DNA

The two sub-methods shown in Fig. 1.5, both use the fluorescence-activated cell sorter (FACS) to select cells based upon the increase in DNA content during the S phase. The left protocol has a higher signal-to-noise ratio by pulse labeling with BrdU (5-bromo2-deoxyuridine), which is a kind of base substitutions that can mark newly synthesized DNA. Then immunoprecipitation technology is used to pull down BrdU labeled DNA (Fig. 1.5 left panel, BrdU-IP method). Because all BrdU-containing sequences are synthesized after the start of the S phase, according to the FACS sort result, the BrdU-containing sequence is divided into two groups, i.e. early S and late S, based on the amount of DNA within cells. On the other hand, the S/G1 method (Fig. 1.5, right panel) is just based on the copy number of DNA got by FACS to classify the cells into G1 and S phase groups, respectively. In G1, replication has not yet started, and cells have only two template parent chains. So, cells in the G1 phase contain equal copy numbers of all genomic sequences. However, since without the BrdU labeling, in S /G1 method, the newly synthesized sequences cannot be distinguished from the template parent sequences, therefore it might generate higher background noise than the BrdU-IP method.

No matter the groups between G1 and S or early and late S phase, after separation of the two groups, people can use a high-density whole-genome oligonucleotide microarray or next generation sequence (NGS) to determine the sequences in each group. For example, by NGS, after mapping the sequences to the genome, by counting the mapped sequence numbers in small interval along the genome (the resolution of interval could be set by researchers), people can estimate the replication timing values as log2(early sequence number/late sequence number) or log2(S sequence number/G1 sequence number), respectively.

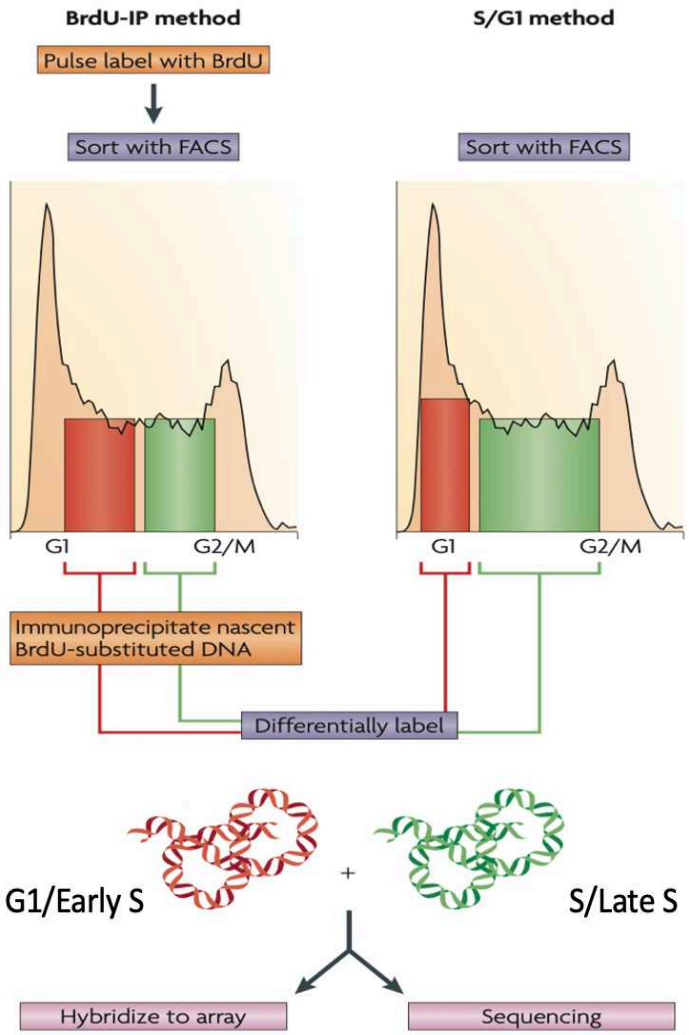


Figure 1.5: Schematic of two classical sub-methods for obtaining log2 Ratio timing values (Gilbert DM et al., 2010). The two profiles with double peaks represent the FACS statical result for cells. The X-axis shows the fluorescent units representing the amount of DNA within a cell and the Y-axis means the number (or proprotion) of cells corresponding to a DNA amount of X value. Red and green boxes are used to mark the two groups in the cell cycle, which are located in sequential timing order. The red one is earlier than the green one. In the right protocol, red indicates the cells in the G1 phase and green is the cells during the S phase. In the left one, all selected cells are from the S phase and will be classified into the early S phase and late S phase.

1.1.5.2 Determination of replication timing by Repli-seq

In the Repli-seq method, similar to log2 ratio methods for replication timing value calculation, the reference genome was firstly cut into 50 kb or 100 kb bins, but it classifies the entire S phase into six S-phase cells subpopulation (Fig. 1.6) instead of two groups like log2 ratio methods. The read numbers from the sequencing data underlying these regions were also applied to the six S-phase cell subpopulations (Fig. 1.7). S50 values is a replication timing estimator that measures the fraction of the S phase at which 50% of a given genomic region has been replicated (Chen et al., 2010; Dellino et al., 2013); The S50 value will be scaled in the range from 0 to 1 indicating very-early to very-late replicating regions, respectively. The smaller the S50 value being, the earlier the

corresponding position will be replicated. Recently, there is a paper to report a similar way to calculate a more precise timing value by high-resolution Repli-seq with 16 fractions of the S phase (Vouzas and Gilbert, 2021).

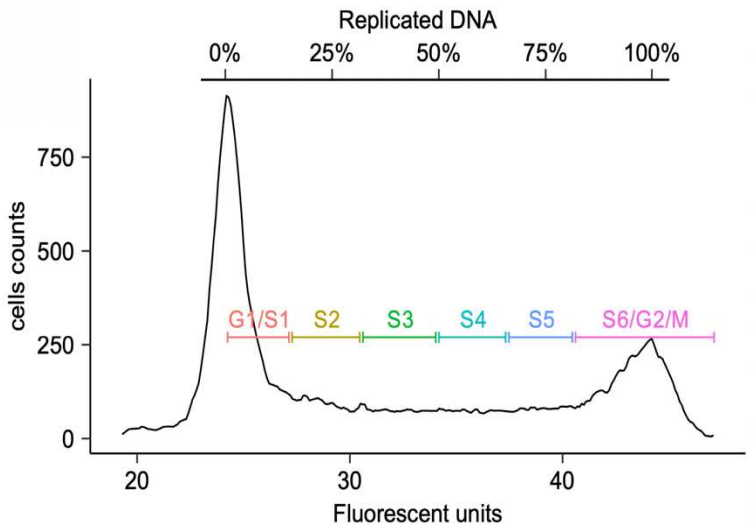


Figure 1.6 Density curve of cell count distribution by flow cytometry (Brison et al., 2020). X-axis shows the fluorescent units representing the amount of DNA within a cell, and Y-axis gives the cell counts for each fluorescent unit position. The flow cytometry is able to classify all the cells into group S1 ~ S6 (from early to late), based on the labeled fluorescent units indicating how much percentage of DNA being replicated in cells.

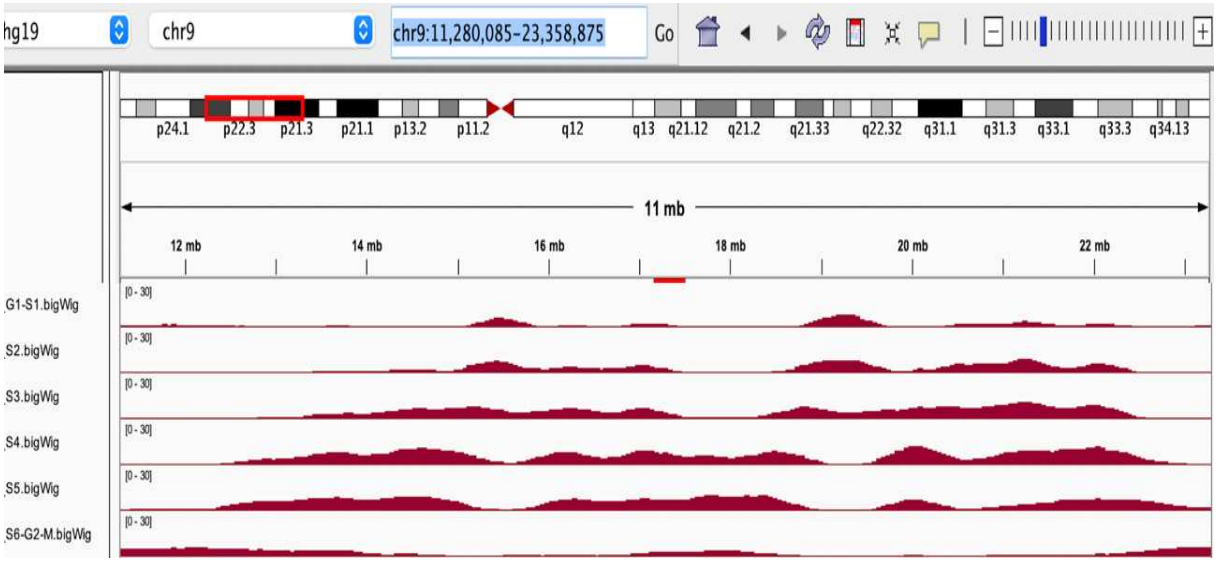


Figure 1.7 Reads distribution of Repli-seq from S1~S6. The IGV screenshot displays the read densities of newly replicated DNA detected in cells obtained in different periods of S phases on chromosome 9, which shows the movement trend of the read distributions from early initiation zones to both sides. In the above plot, the replication initiation zones can be roughly observed in the figure, which is located in the positions enriched BrdU labeling sequences of the G1 phase, but since the resolution of Repli-seq is in 50 kb or 100 kb, it is not enough to meet the requirements for precise positioning of the replication origins.

1.1.5.3 T-peak regions containing replication initiation zones

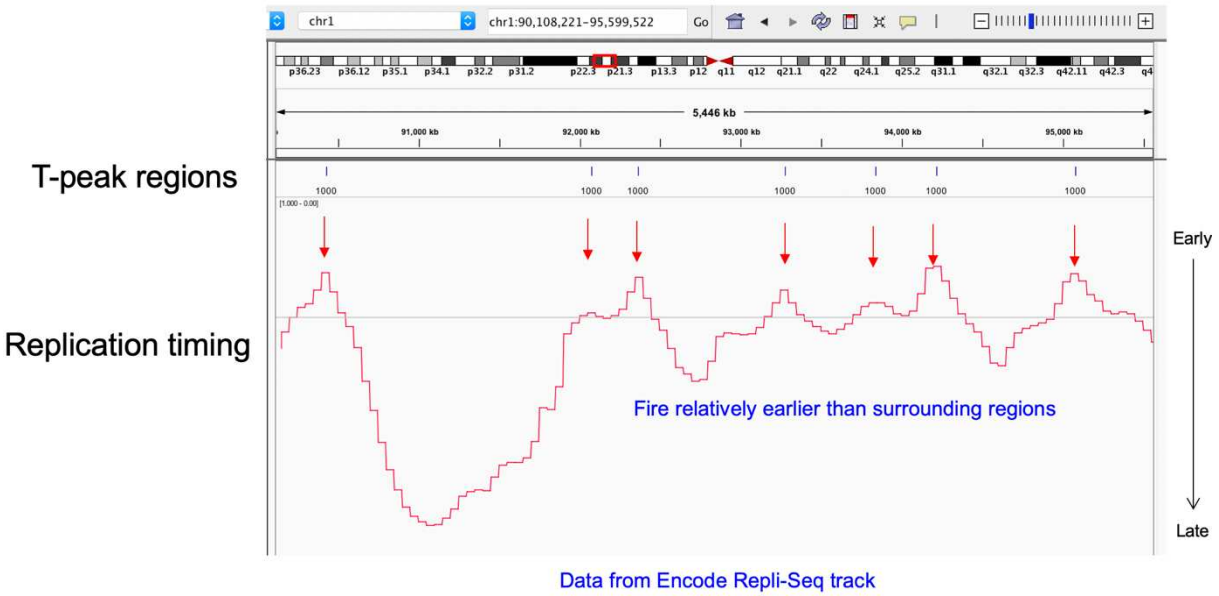


Figure 1.8 The schematic of T-peak regions by IGV. The first line gives the T-peak center regions marked by blue bars with 1 kb width. And the red step curve is the timing curve composed of the replication timing values in 50 kb non-overlapping bins. The right-side black arrow reveals the relationship between the height of the timing value and the tempol order of replication. The higher the timing value of a given position is, the given position will replicate earlier. And all T-peak regions are consistent with the positions where replication timing curves up to the local maximum values.

For any specific cell line, people can draw replication timing curves in different resolutions (e.g. different bin sizes, although the resolution of Repli-seq is limited by labelling time and/or sequence depth) along the genome. Hence, we can easily get the replication temporal order in any partial area. According to the replication timing curves, we take the positions, the replication timing of which is earlier than the neighbored regions, to define the replication timing peaks, called T-peaks (Fig. 1.8), which should contain replication initiation zones.

1.2 Replication regulation in timing and origin location

1.2.1 The genetic and epigenetic modifications around origins

DNA replication process organized by licensing, Pre-IC formation, and firing steps (Fig. 1.3). The DNA replication must occur in accessible regions with unwinding DNA single strand. So, the early firing origins are frequently located at open chromatin regions and are highly associated with the epigenetic modifications related to several open histone markers, such as H3K4me1, H3K79me2, pho-RNA Pol2. DNase I digestion is a commonly used method to mark the open chromatin regions. At the same time, the DNA replication also needs various proteins to help to fire the replication origins. For example, H2AZ is recently recognized as a factor that can facilitate licensing and activation of early replication origins (Long et al., 2020). Some studies have also reported that replication origins may correlate with transcription start sites (TSSs), CpG islands (CGIs), and G-quadruplex (G4) sequence motifs (Karnani et al., 2009; Masai et al., 2010; Mukhopadhyay et al., 2014).

1.2.2 Stochastic model of initiation-timing regulation

As for whether a given origin will be fired and when it will be replicated, currently, there are several models to explain, and they are still under debated. The main models are the Domino-like model (Sporbert et al., 2002), deterministic model (Lebofsky et al., 2006) and stochastic model (Rhind, 2006).

Domino-like model suggests that the origin firing is triggered by replication of adjacent regions in a next-in-line mode. There is a paper (Guilbaud et al., 2011) reporting chromosomal regions in HeLa cells with sequentially activated origins that are neither clearly early nor clearly late replicating. In addition, concerned with the chromatin folding, such adjacent effect could even cause spatial effect to amplify the 1D replication cluster along DNA sequence to 3D replication cluster in heterochromatin during late S-phase (Löb et al., 2016).

Besides that, deterministic models suppose that different initiation sites are programed to initiate at different, well-defined times. Stochastic model posits that different initiation sites have different initiation probabilities but can fire at any time during S phase (Rhind et al., 2010). Furthermore, some papers propose the combination of stochastic and deterministic models (Labit et al., 2008).Whether metazoan initiation timing is stochastic or deterministic, or some combination of the two, is still very much an open question (Bechhoefer and Rhind, 2012).

1.3 The current technologies used for origin identification by bulk data

1.3.1 SNS-seq

Two essential derivatives produced during DNA replication are short nascent strand (SNS) DNA and Okazaki fragments. Some studies are intended to trace back the firing origins by locating the SNS. It should be noted that, whether it is a short nascent strand or an Okazaki fragment, they have an RNA primer at the 3' end (Fig. 1.2), which can protect their sequence from exonuclease hydrolysis from 5' to 3'. Therefore, the DNA of the asynchronous cells containing S phase cells is extracted, DNA will be purified with λ-exonuclease digestion to remove all the contaminant SSS (short single-strands) due to sheared DNA. Finally, only SNS and Okazaki fragments are left. The average length of a short nascent strand is around 1~1.5kb and Okazaki fragments are with a length from 150~200 bases. So, by size selection, it is easy to separate them and map short nascent strands to genome reference by next generation sequencing. Then the piled-up signal peaks of the isolated short nascent strands in population-based data provide the replication origin positions (Fig. 1.9). People call this method SNS-seq.

Currently, SNS-seq has been applied to mice, *Drosophila* (Lombraña et al., 2016), and different human cell lines, such as Hela, IMR90, IPS, H9, etc. (Picard et al., 2014). More than 200,000 potential initiation sites on the human genome are founded by SNS-seq (Besnard et al., 2012). These origin sites are size-specific origins with an average length of 760 bp and cover 6% of the human genome (Picard et al., 2014). Some studies also show that active origin sites often correlate with transcription start sites and are located in GC-rich regions, near CpG islands and G4 (G-quadruplex secondary structures) sequence motifs (Besnard et al., 2012; Langley et al., 2016).

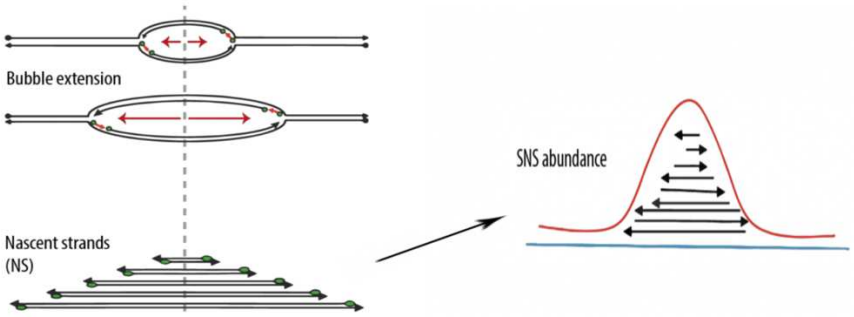


Figure 1.9 The illustration of SNS-seq principle (Francesco De Carli, 2017). The upper left part shows the bubble extension process. There will be replication bubbles of different lengths in asynchronous data. The long black arrows representing nascent strands and the small red arrows on the lagging strand representing the Okazaki fragments. The bottom left part is the isolated short nascent strand from replication bubble of different sizes with sorted length varied from 1~1.5 kb. The green dots on the plot are the RNA primers, which can protect only the SNS sequence and Okazaki fragments from the digestion of exonuclease. After λ-exonuclease purification treatment, the positions with SNS (>1 kb) enrichment will reflect the replication origin positions. The right side shows the enriched SNS signal peak around a replication origin site.

Based on SNS-seq, Mukhopadhyay and colleagues have developed BrIP-SNSseq (Mukhopadhyay et al., 2014), a variant of the technique consisting of sorting the BrdU-labeled DNA at increasing time-points during the S phase. The relative amount of SNS at each point in the different fractions allows computing the genome-wide replication timing profile, too. Meanwhile, the introduction of BrdU can also further reduce the noise data from not fully digested short single-strands. However, some studies point out the result of SNS-seq may contain a lot of false positive data, because of the hard control of exonuclease's activity. A comparable number of ORI (origin replication initiation site) peaks was obtained with or without λ-exo treatment (Valenzuela et al., 2011). It was also proposed that the strength ORIs might have been overestimated about tenfold, considering that an accumulation of small (~200 bp), duplex DNA molecules (proposed to represent abortive initiation intermediates), was detected in total genomic DNA (Gómez and Antequera, 2008). Even only 56.5% of the BrdU-SNS-seq peaks accorded with 50.2% of these SNS-seq peaks (Picard et al., 2014). This poor consistency raised a debate on whether it is possible or not to detect real SNS with prior λ-exo digestion, considering that the vast of "SSS" (short single-strands) DNA is still mixed up.

At the same time, another controversial topic is the relation of replication origins and G4. Some studies report the origins detected by SNS-seq are associated with G-quadruplex consensus motifs (Besnard et al., 2012) and suggesting G4 as an potential regulator of origin function. But some other studies propose that this association may be due to the fork stalling. Federico and colleagues have found the role of origin-proximal G-quadruplexes which tend to stall replication forks *in vivo* transiently (Comoglio et al., 2015). The short nascent strands following the replication forks also stop around the G4, which takes the stalling position as origins located in G4 regions. Therefore, if it is the case, the final SNS-seq result will be mixed with a part of these false positive signals.

1.3.2 Bubble track

Bubble track uses the special structure of replication origins to hunt for their positions. When the replication process starts to fire at a replication origin, a bubble structure will form with two divergent replication forks. Due to the read length limitation of next generation sequencing technology which is only 100~500 bp, the longer DNA molecules had to be broken into smaller fragments to get sequence information, and then based on the overlapped part of read information to assemble complete sequence information. The shotgun method (Weber and Myers, 1997) is the technology to cut the DNA randomly into small segments. The traditional shotgun uses restriction endonucleases to cut the recognition sites on the target DNA to form short sequence fragments that follow a normal distribution within a certain length range. Besides that, the other common method uses ultrasonic DNA fragmentation to break large molecular DNA into small fragments of about 350 bp. In bubble-seq, after shotgun treatment with restriction endonuclease the ~2 kb DNA fragments obtained could be 3 different shapes, i.e. the linear DNA, the Y-shaped replication forks, and O-shaped replication bubble. However, these 3 DNA fragments have different speeds under agarose gel electrophoresis. At the molecular level, agarose has many dendritic structures. The linear DNA can pass polymerizing agarose fibers relatively easily, followed by Y-shaped replication forks, and replication origins could be hung on the dendritic structure with the slowest speed. In this way, the replication bubble could be isolated and detect the origin positions by next generation sequencing. The bubble traps were validated by 2D gel electrophoresis, which confirmed a purity with >80% replication bubbles (Mesner et al., 2011). Bubble track (or Bubble-seq) has found more than 50,000 initial zones cluster by origins detected by bubble track in human GM06990 cell line along the genome (Mesner et al., 2013), and also applied for Hela cell line to generate the library within ENCODE pilot regions covering ~1% of the human genome (Mesner et al., 2011). In GM06990 cell line, the average and the median zone size are 20 kb and 16 kb, respectively. Around 32% of initial zones are early-firing with the highest origin density. The late initial zones in the 1 Mb scale have the 17% lower than early-firing initial zones in origin density, followed by initial zones in the mid-S-phase with the lowest origin density. However, only 45-46% of the SNS-seq peaks overlapped the 36-37% of the bubble-seq (Picard et al., 2014). This may be due to the reason that bubble traps identified large initiation zones that are variable between cell lines, while SNS called sharp initiation peaks that are more conserved between cell lines. Besides that, bubble track is also limited by the relatively large sizes of Y-shape fragments with slow-moving speed, too, which can cause false-positive noise results.

1.3.3 MCM / ORC ChIP-seq

As shown in Figure 1.3, origin licensing is the first step of DNA replication initiation. The origin replication complex (ORC) is an essential element of Pre-RC to finish the complete replication initiation process. But even if the licensing origins finished, they may still keep silent, and not continue the following Pre-IC formation and origin firing process. Therefore, the genomic regions bound by ORC may not be the final firing origin sites. However, no matter what, it was still proposed very early as a method to find all potential origin sites, and ORC ChIP-seq (Coupling chromatin immunoprecipitation) came into being. This method captures the target protein ORC by immunoprecipitation and then detects all licensing positions binding to ORC by next generation sequence.

In MCM ChIP or ORC ChIP-seq, in order to reduce technical noise, the researchers openly performed multi experiments, i.e. several biological replicates, in unsynchronized cells to get the sequences binding with MCM or ORC. Then based on the overlapped regions between several replicates, or within 0.5 kb inter distance as criterion, to pick out proper regions showing up in all or most replicates as potential replication regions (Fig. 1.10). Then using SNS-seq and DNase-seq data as supporting evidence to classify the potential origins into firing origins or dormant origins. The origins close (within 0.5 kb) or overlapped with SNS-seq regions and show DNase-seq signal peaks will be recognized as firing origins, and the other regions will be more probably dormant origins. There are around 200,000 MCM7 peaks were found in several replicates of Hela cell line and 78,257 sites are associated with SNS-seq origins (Sugimoto et al., 2018). In the latest research of human cells, people found the distribution of ORC and MCM is dependent on transcription and depleted from transcribed gene bodies. But they are enriched in the TSSs (transcription start sites). ORC/MCM genomic distribution has an obvious correlation between replication timing but not related with initiation zone (Kirstein et al., 2021).

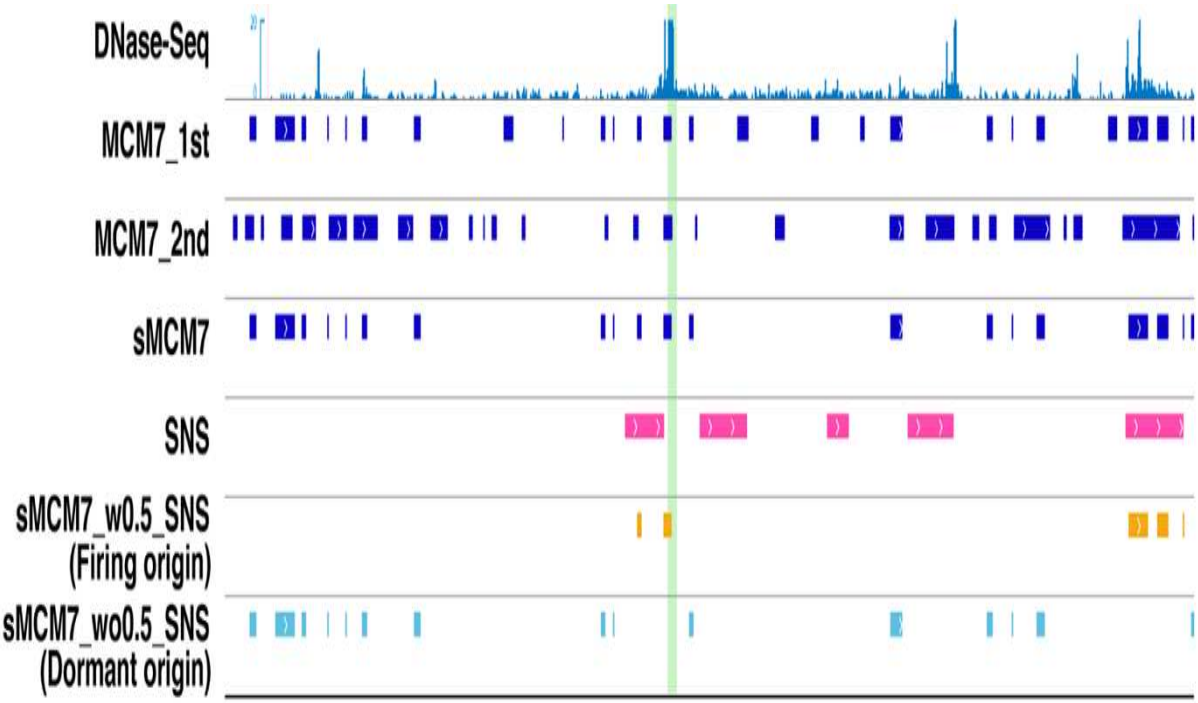


Figure 1.10 Genome-wide identification of firing and dormant origin sites (Masatoshi Fujita1 et al., 2018). First line DNase-seq, DNA footprints by DNase I digestion. The second and third lines (MCM7_1st and MCM7_2nd) are the two replicates of immunoprecipitation results for MCM7, and sMCM7 is the filtered overlapped potential origins after comparing MCM7_1st to MCM7_2nd. The 5th line is reference SNS-seq regions. The last 2 lines are classified as firing origins and dormant origins based on whether it overlaps with DNase1 signal peaks marked by green column or overlapped with SNS-seq regions.

Now the MCM/ORC ChIP-seq has applied to yeast, fruit flies, and human cells (Dellino et al 2013, MacAlpine et al 2010, Miotto et al 2016, Xu et al 2006). In *S. cerevisiae*, Autonomously Replicating Sequences (ARSs) contain a consensus sequence (ACS) that can be bound by ORC has been found essential for origin function. In addition, there is near perfect concordance between ORC and Mcm2-7 binding peaks. But in *Drosophila*, there is a vast excess of Mcm2-7 relative to ORC assembled onto chromatin when cyclin E/CDK2 activity rises in late G1 (Nina Kirstein et al

2021). These excess Mcm2-7 complexes exhibit little co-localization with ORC or replication foci (Sara K Powell et al, 2015). In humans, the Epstein–Barr virus (EBV) was used, whose replication in latency is entirely dependent on the human licensing machinery, to compare ORC and MCM binding and replication initiation sites. It has been shown that, there are a five- to tenfold excess of potential origins are licensed per genome with respect to 1–3 mapped initiation event, which means human replication initiates in zones, which comprise multiple, individually inefficient sites (Kirstein et al., 2021). Besides that, ORC has many functions other than DNA replication. Some ORC proteins work as transcription factors as well (Chesnokov, 2007). So, the detected origins of ORC-ChIP may be related to other functional genomic regions than replication origins. For example, ORC ChIP-seq identified 13,600 ORC1 binding sites in human HeLa cells, which do not reveal any sequence consensus (Dellino et al., 2013). Only 11-30% of these peaks overlapped SNS-seq peaks, and 47% overlapped bubbles. All of these biological reasons result in a huge controversial result in ORC ChIP-seq technology.

1.3.4 EdU-seq-HU

The incorporation of halogenated nucleotides is a conventional method to monitor the ongoing replicated regions during the S phase. The BrdU, which is frequently used in replication studies, can be detected by anti-BrdU antibody only after the DNA becomes single-stranded due to resection (Mukherjee et al., 2015). EdU (5-ethynyl-20-deoxyuridine) is another thymidine analog that has some technical advantages over BrdU usage, since EdU will be conjugated to fluorescent aside by Cu(I)-catalyzed reaction and can be detected in double-stranded DNA (Hua and Kearsey, 2011). Unfortunately, the EdU is toxic to the cells and activated the rad3-dependent checkpoint, which likely blocks over mitosis. Toxicity effects of EdU on mammalian cells have also been reported, suggesting that EdU may not be suitable for continuous labeling studies (Hua and Kearsey, 2011). So, it often takes several times to confirm the mark position by EdU in more than one cell cycle (Diermeier-Daucher et al., 2009; Hua and Kearsey, 2011). In addition, HU (hydroxyurea) can arrest fork progression after origin firing. Therefore, under the HU treatment on cells synchronized at the S phase entry will allow enriching the EdU signals around the replication origins. For a limited DNA synthesis situation, in the hydroxyurea-treated cells, EdU incorporation can be easily detected under fluorescence microscopy. Thus EdU-seq-HU protocol has been developed to locate the early replicated origin positions (Macheret and Halazonetis, 2019). The problem is the cell arrest led to an incomplete cell cycle, which can only detect the origins fired at the beginning of the S phase, and cannot identify the replication origins fired in other periods, such mid or late S phase.

1.3.5 Ini-seq

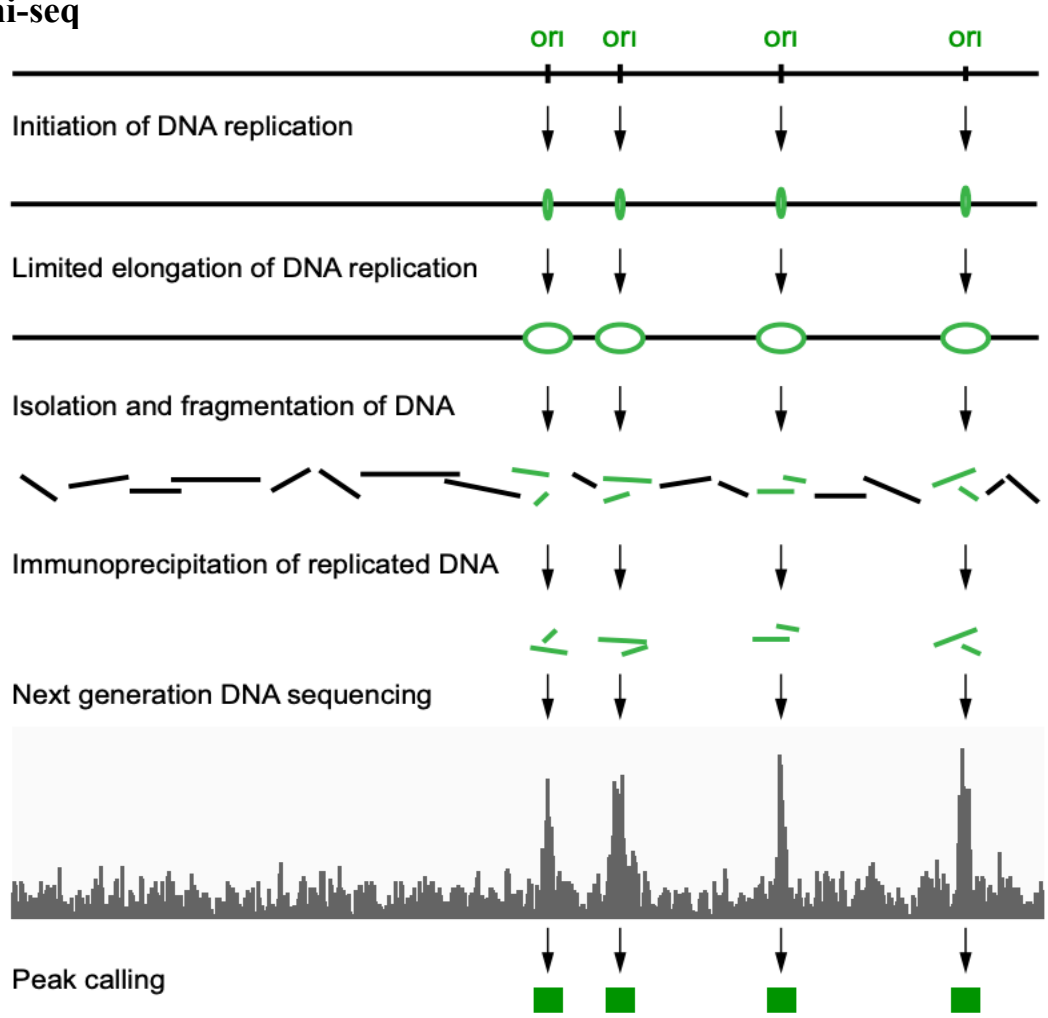


Figure 1.11 Schematic of ini-seq experiment protocol (Langley et al., 2016). Green lines represent digoxigeninlabeled nascent replicated DNA labeled by dUTP, and black lines represent unreplicated double-stranded DNA.

Another approach for replication origins hunting is ini-seq (Fig. 1.11). This is a method based on an *in vitro* system and applied to human EJ30 and Hela cell lines. Firstly, the nuclei have been extracted from human synchronized cells in the G1 phase. By adding an extract of proliferating cells makes replication start, and the newly synthesized DNA will be labeled by dUTP. Then the immunoprecipitation will be used to pull down the newly synthesized DNA labeled with dUTP. Sequencing, mapping, and peak calling allow identifying specific replication origin sites along the human genome. The median length of the origin sites is 1,184 bp.

1.3.6 OK-seq

Okazaki fragments must be generated on the lagging replicating strand of opposite replication forks during the replication process at the two sides of origin positions (Fig. 1.12). In another word, if we can detect the distribution of piled-up Okazaki fragments, we can detect the replication origins. OK-seq introduces a novel conception called Replication Fork Directonality (RFD): indicating the proportions of rightward- (R) and leftward- (L) Okazaki fragments at each position along the genome, like the formula, shown in Fig 1.12: RFD = (C-W)/(C+W), where C and W correspond to the numbers of detected Okazaki fragments mapped on Crick and Watson strand, respectively. Since we can calculate the RFD values at each position of the genome, the RFD curves along the entire genome can be drawn and origin regions corresponding to increased RFD curves shift can be determined as indicated in Figures 1.12 & 1.13.

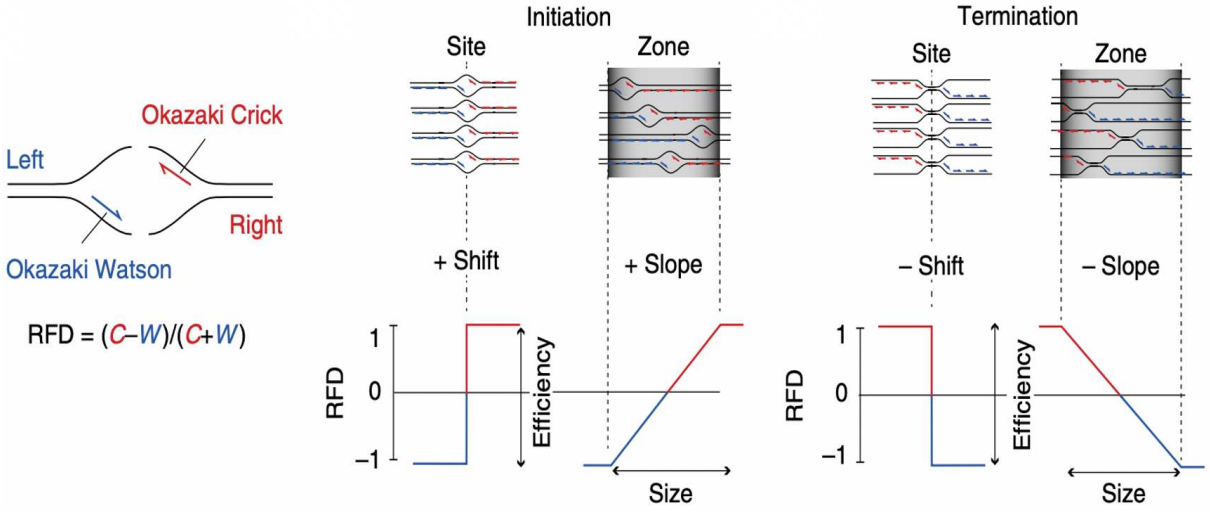


Figure 1.12 OK-seq principle (Petryk et al., 2016). The left panel shows the principle of OK-seq and the formula in calculating Replication Fork Directionality (RFD) based on the OK-seq data. Based on the strand where Okazaki fragments come from, we can classify all of them by left and right forward replication forks that they belong to. The blue Okazaki fragment on the lagging strand of left moving replication forks is named Okazaki Watson, and the red Okazaki fragment on right moving replication forks named Okazaki Crick. After mapping bulk Okazaki fragments to the whole genome, we can count the number of Okazaki Watson (C) and Okazaki Crick (W) in each position and calculate the RFD = (C-W)/(C+W) along the entire genome. In an ideal case, if all cells select the same origin position to fire, the RFD curve will become a vertical ascending step from RFD = -1 on the left side of origin to RFD = 1 on the right side of the origin. In the case with different origin selection within an initiation zone in different cells, normally, the RFD curve around the origins should be like an increasing slope shape. Similarly, in the termination position of a fixed termination site and a termination zones, the RFD curve will show a descending shift and decreasing slope, respectively.

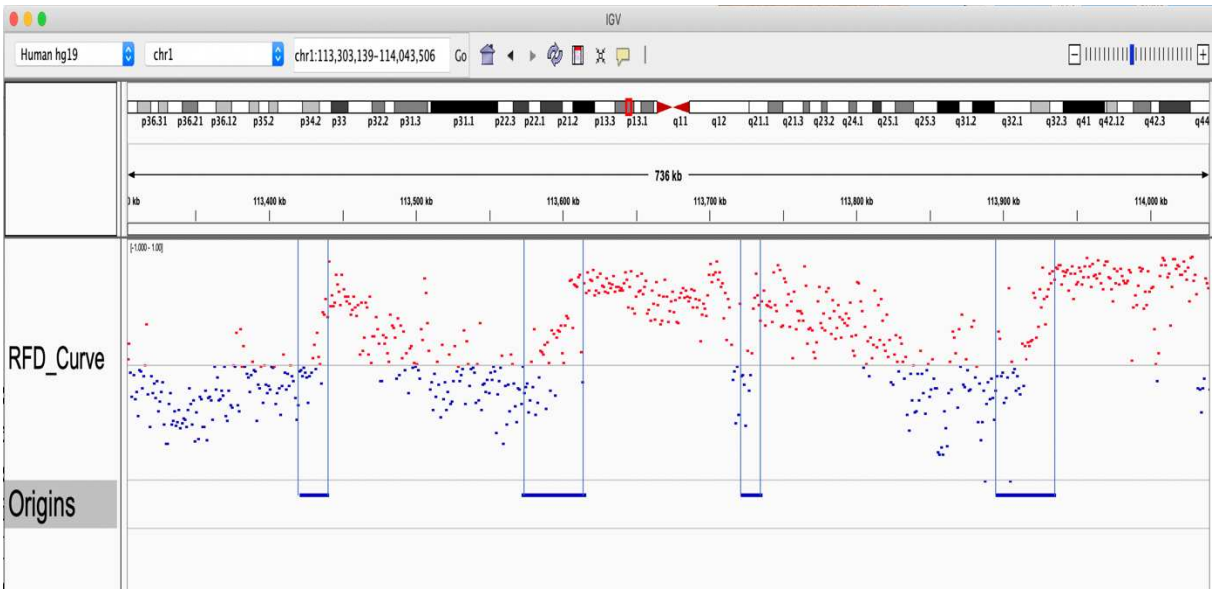


Figure 1.13 OK-seq replication origins called by RFD curve. Each dot indicates the RFD calculated within each 1 kb non-overlapping bin. The vertical blue lines at the positions of two ends of origins show the increased shift trend from negative (blue) to positive (red) RFD values.

The protocol of the experiment based on the above theory could be briefly summarized as 5 steps.

(1) EdU/EdC labeling (1~2 mins) marks the newly synthesized DNA including Okazaki fragments.
(2) Genomic DNA extraction by standard proteinase;
(3) Okazaki fragment isolation and biotinylation: centrifugation can isolate the <200 nt single-stranded DNA, and then biotin-TEG-Azide can pull down the EdU labeled Okazaki fragments.
(4) 2 pairs of adaptors ligation for purified Okazaki fragments, which permit the mutual authentication; and all Okazaki fragments will be captured with 200 mg of Dynabeads MyOne Streptavidin T1 according to the manufacturer's protocol.
(5) Classical next generation sequence protocol including library amplification by PCR, sequencing data, and alignment to the genome.

Tracing the history of OK-seq, this method has been applied in yeast and humans successively through continuous optimization and evolution, 1st OK-seq (with a different experimental design) was performed in yeast in a ligase mutant (Duncan J. Smith et al, 2012), then it succeeds on the WT human cells by sequencing the highly purified short (<200 nt) EdU (or EdC) labeled single-stranded DNA, which highly enrich Okazaki fragments (Petryk et al., 2016). By OK-seq, Petryk and colleagues have shown that replication initiates stochastically in human cells, primarily within non-transcribed, broad (up to 150 kb) initiation zones that often abut transcribed genes and terminates dispersively between them.

1.4 The current single-molecule technologies used for origin identification

All the methods discussed before are origins or potential origins detection from various population-based data. There is a low agreement amongst various genome-wide studies. Regardless of the mechanism level, the major debate is whether replication origins are located at specific sites or stochastic occurred in broad initiation zones. And most methods, more or less, have their own technical or biological problems, leading to different population-based methods that might identify different "types" of origins. Whatever the main reason for the controversial results is the heterogeneity of the choice of replication initiation between cells. At present, the best way to solve this problem is to detect the origin of replication at the single-molecule level. Below I will introduce several commonly used single-molecule detection methods.

1.4.1 DNA combing

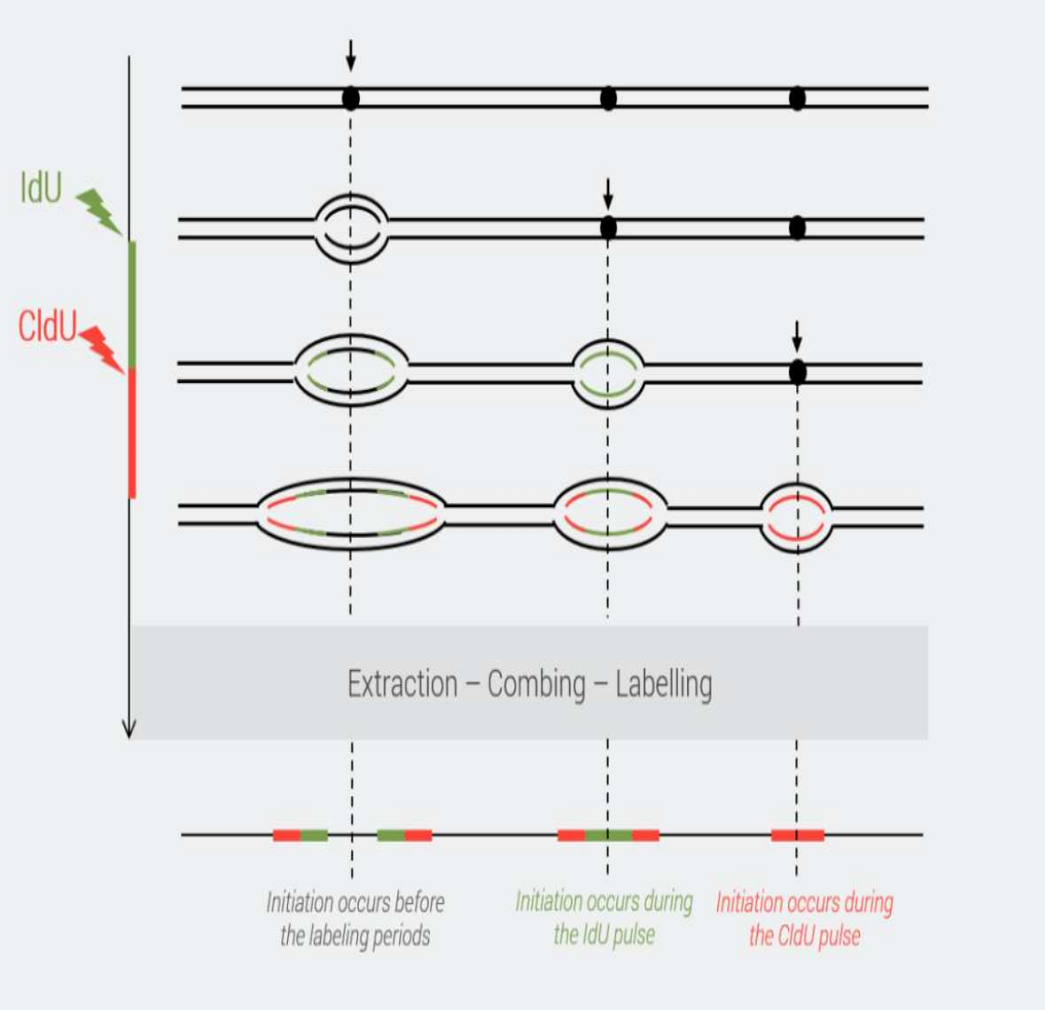


Figure 1.14 Schematic of DNA combing experiment process (from the introduction of genomic vision company). Two different fluorescent dyes to label the newly synthesized DNA sequentially. The green lines represent IdU with green fluorescence dye and red lines represent CIdU with red fluorescent dye.

DNA combing is the first single-molecule method applied to replication site detection invented by Bensimon (Allemand et al., 1997). DNA combing marks the newly synthesized DNA by two thymine analogs, such as IdU and CIdU, sequentially (Fig. 1.14). Then, DNA extraction will be performed to isolate intertwined DNA, and extracted DNA will move to the surface of a vinyl silane treated glass carrier. At the end of DNA molecules, there is a specific pH26 exposing polar groups which can bind to ionizable groups coating the hydrophobic surface. Meanwhile, the mid-segments show weaker negative than DNA ends, so, only DNA ends bind to the silanated coverslip. This will follow by stretching labeled DNA molecules to a linear structure by capillary force between two glass coverslips. Moreover, performing FISH (Fluorescent in situ hybridization) on combed molecules permits their genomic identification (Tuduri et al., 2010), although it's technically challenging.

With the help of intermittent analysis for fluorescence colors by high-resolution microscope, it can clearly indicate the location of replication origins on individual DNA fibers (Fig. 1.14). Compared with the methods introduced before, the biggest advantage of DNA combing is that it is a single-molecule technology. The other methods are peak calling result from bulk data, which exclude the location with weak signals. The origins obtained by the population methods often enrich in the early replication initiation sites shared by multiple cells. And for those replication initiation sites that only fire in late S phase or origins that only fire when replication fork stalling occurs, they can only be detected by method at the single molecular (SM) level. DNA combing is one of such SM approaches. All detected origin locations are real origins. Nevertheless, the limitation of this methods is very low throughput and lack of sequence level resolution. This disadvantage makes DNA combing can't apply for the genome-wide origin detection.

1.4.2 Nanopore sequencing

Nanopore sequencing is a technology that can detect the sequence by different resistance of bases when the DNA sequence passes through the magnetic beads with electrodes. The magnetic beads continue to discharge, and as the sequence continues to enter the magnetic beads, the base sequence passing through the magnetic beads is continuously replaced. The base sequence of different resistances will cause the current signal to change, so as to distinguish the four bases of A, T, G, C. Similarly, based on the current signal difference, *in vitro*, it is able to distinguish newly replicated DNA marked by BrdU (5-bromo2-deoxyuridine) with normal base dTTP in unreplicated regions like Fig. 1.15 (Müller et al., 2019). There is a nanopore electric signal feature to detect the BrdU by machine learning. *In vivo*, it will be barely qualified for thymidine detection from BrdU by model optimization, but there will be false positive BrdU phenomenon occurred in the BrdU-enriched sample reads (Hennion et al., 2018).

A nanopore is indeed a high-throughput detection method compared to DNA combing at the single molecular level. Moreover, compared with the non-single-molecule method, because DNA fiber does not need to be cut by the shotgun method, the ultra-long sequence is directly analyzed. Now, nanopore sequencing is successfully applied to the yeast at near-nucleotide resolution (~200 nt) and found 58,651 replication tracks (Hennion et al., 2020). The sequencing read length is tens of kb or even 100 kb (with few molecules), in yeast, the read length is between 10~140 kb for the BrdU labeling sequence and similar for normal dTTP sequence (Hennion et al., 2020). Although

the matching accuracy is far beyond the next generation sequencing technology, it can be used for matching with multi repetitive sequences region with the long reads, this is not possible with second-generation sequencing. However, the high sequencing cost limits the sequencing coverage, therefore, hard to apply to study DNA replication of human cells genome-wide.

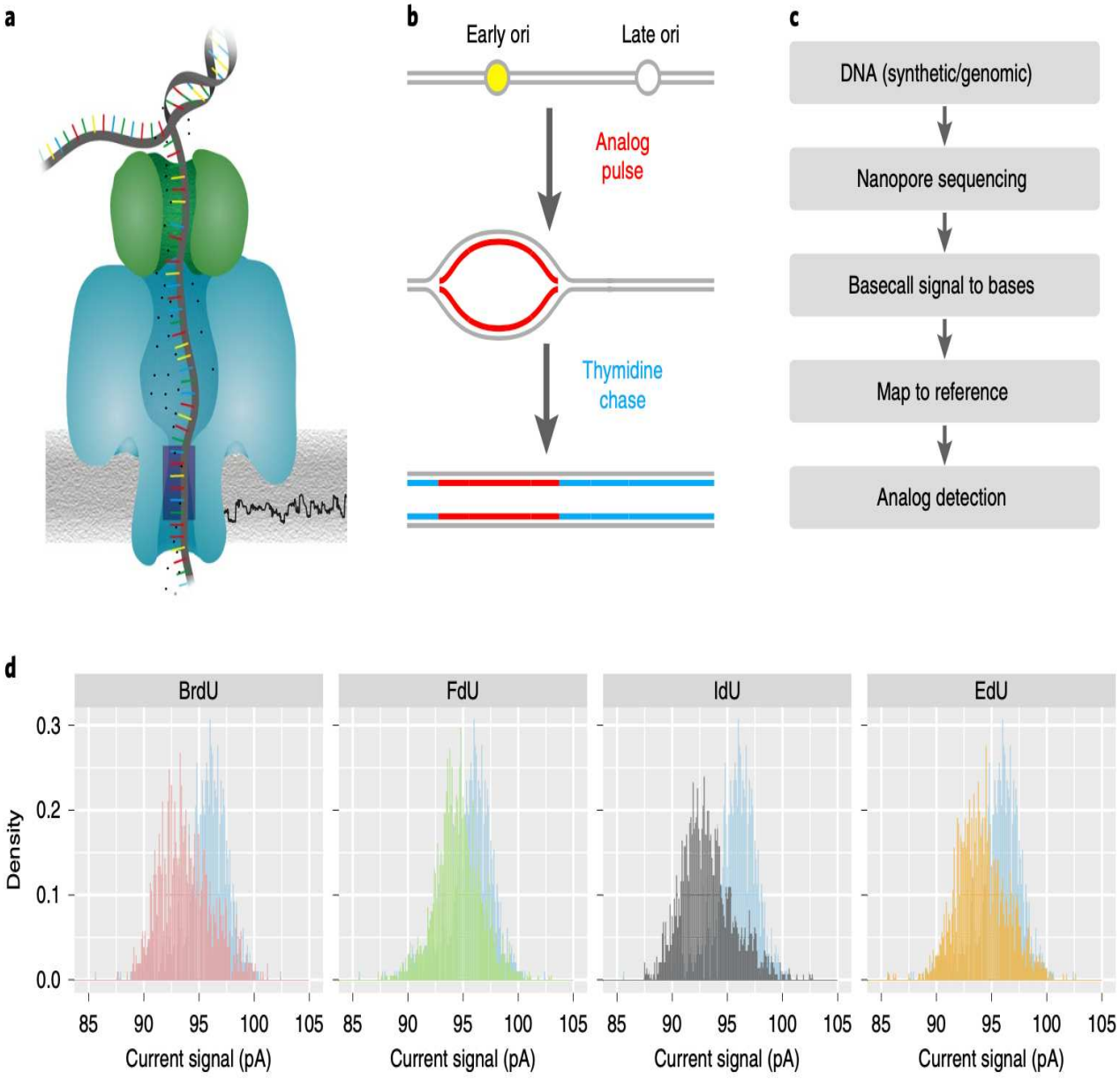


Figure 1.15 Schematic of Nanopore technology (Müller et al., 2019)**. (a)** The technical principle of Nanopore. With the one base step length, DNA passes through the magnetic beads with electrodes little by little and calculates the resistance of the DNA fragments wrapped in the magnetic beads through the applied voltage and current, each time they move. Determine the sequence information according to the different resistance values of the four bases of A, T, G, and C. **(b)** Schema represents the origin detection principle. Like several methods introduced before, Nanopore also used BrdU or another thymidine analog to mark ongoing replication regions like the red line in plot b, and the blue line is the normal DNA sequence. **(c)** pipeline for the origin detection process by Nanopore. **(D)** The current signal distribution of thymidine (blue) compares with distributions of 4 different thymidine analogs.

1.5 A novel method: ORM (optical replication mapping)

A variety of methods of replication origin detection have been listed above. More or less, these methods have their own shortcomings in biological or technical means. For a population-based approach, SNS-seq may accumulate the short nascent strand close to the G4 region due to fork stalling, which causes false positive origins. EdU-HU is toxic to cells, which makes continuous labeling become hard and may have an effect on the physiological process of cells. But currently, the biggest problem for EdU-HU is that it can only detect the replication origins in the early S phase after synchronization by HU. Similarly, when it comes to the impact on the replication process, nuclear extraction in ini-seq also has the same problem. It is unknown whether the *in vitro* system constructed by ini-seq can 100% simulate the DNA replication process *in vivo*. Concerned with the versatility of ORC, ORC-ChIP-seq contains false-positive results that may be more related to transcription instead of replication. OK-seq may not have biological bias, but the major problem of OK-seq is that if you have a transition, you can identify the initiation zones, while it does not mean that all origins/initiation zones can generate upward transitions. For example, in late replicating regions, the initiation is more or less random, the RFD is close to 0 (with equal probability replicated by leftward or rightward replication forks) within these regions. Not to mention, there is only a very limited consistency between these results. Due to the low fire efficiency of replication origins and the heterogeneity of origin selection. Undoubtedly, the single-molecule method with high sensitivity is the potential way to solve these problems. But DNA combing doesn't have enough throughput to support genome-wide detection. Nanopore-seq can't be applied to the human genome because of the high cost. In summary, nowadays, the main requirements for detecting replication origins are high-throughput, at the single-molecule level, ultra-long DNA molecules for precise alignment, also take into account the cost and the coverage. Therefore, a new optical matching method has emerged, which takes all of the above advantages.

1.5.1 Bionano high-throughput DNA fiber mapping

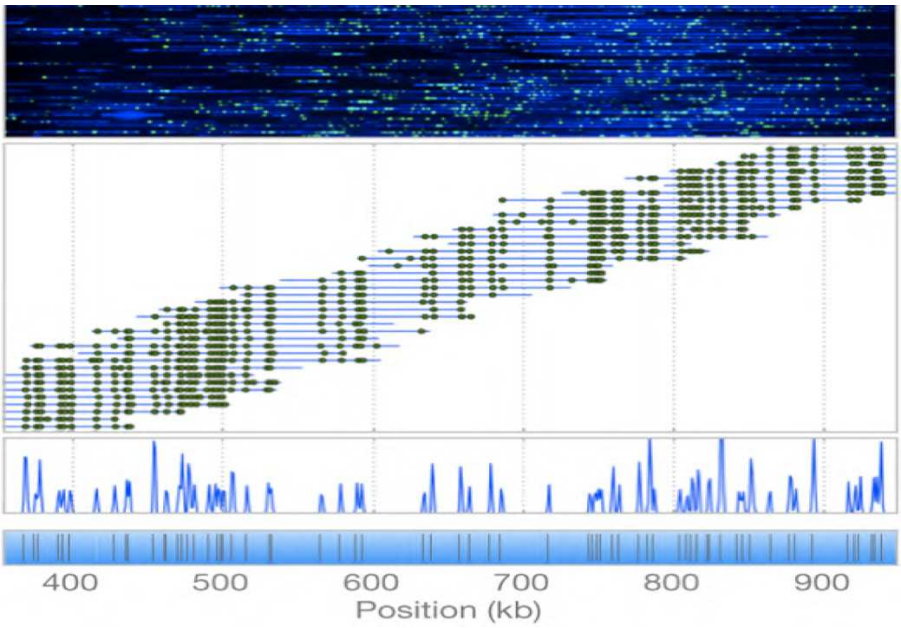


Figure 1.16 Schematic of Bionano principle (from Introduction of the Bionano Genomics company). The blue line labels DNA fibers and green dots on the blue line is the green fluorophore (Nt.BspQI sites or DLE1 sites) that recognizes a specific motif sequence. The top photo is the raw picture, and the bottom shows the alignment of molecules to the reference genome.

Originally, Bionano was a technology used for de novo genome assemblies. For some conserved motifs that repeatedly appear on the genome, tag such motif with fluorophores. Because the reference sequence is known, the position of the fluorophore corresponding to the motifs on the genome is also determined. When DNA fibers are also labeled by the fluorophore, researchers can map the DNA fibers to the reference according to the relative position of the fluorophore on the fibers (Fig. 1.16).

The Bionano platform uses electrophoresis to controls the movement of DNA from the flowcell. The upstream micro-and nano-structures gradience can gently unwind and guides DNA into the NanoChannels. Only stretched linear DNA fibers are allowed to flow through NanoChannels and a high-resolution camera will image them once DNA molecules enter into NanoChannels. In addition to YOYO-1 label DNA molecules (in blue), the Bionano platform also equipped two additional channels for detecting two kinds of color signals (i.e. green and red). One is the green fluorophore used for mapping sequence to reference. Recently, the red channel has been applied to origin detection (De Carli et al., 2018). In a similar manner, in our optical method, we used red dUTP signals for labeling ongoing replication regions or replication origins in synchronized cells (Fig. 1.17).

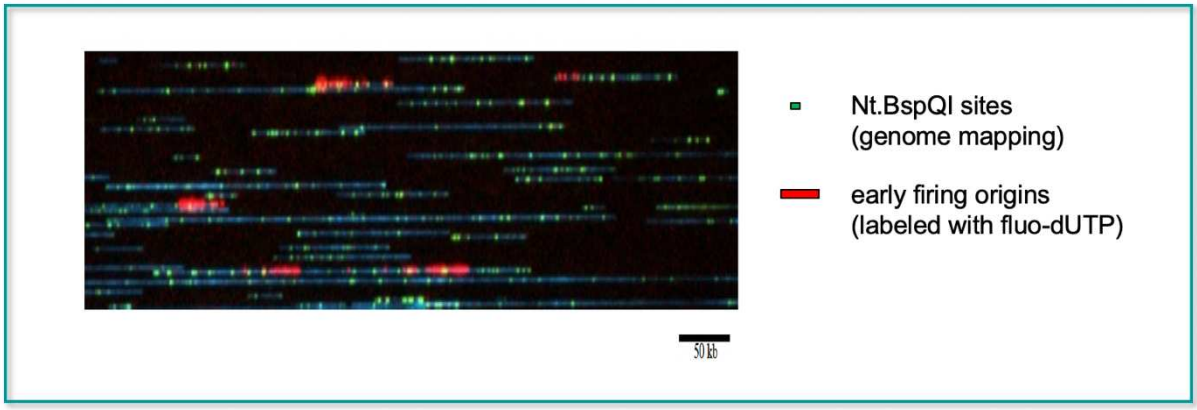


Figure 1.17 Physical image of ORM methods. The red regions are clustered by sparse dUTP signals. Just because they are too close together forming a piece, which looks like a red area. But each red signal has a certain genomic position during the data process. Blue lines and green dots are DNA fibers and mapping green fluorophore (NLRS sites or DLE1 sites)

The average length of DNA fibers analyzed by Bionano can be up to around 300 kb and coverage can be up to 300x of the human genome with one run of Bionano imaging of the latest Bionano system. So, it not only meets the requirement for high-resolution ultra-long fibers like nanopore but also with such a qualified coverage at the single molecular level. Like all single-molecule methods, it can detect the initiation events with low fire efficiency. This is impossible for methods of bulk data. We can have more comprehensive origin information to study the DNA replication process.

Furthermore, for any given position, we can calculate the ORM signal density as fire efficiency to describe the probability of initiation occurred in this position. The accessibility of fire efficiency calculation is a huge advantage for Bionano over other approaches to detect the replication origins,

because there is so much information that can be further mined and analyzed based on it. Firstly, based on the fire efficiency in different replicates, we can observe whether there are independent, fixed origin sites that show high fire efficiency in all replicates. If not, how initiation event occurred, follow the stochastic model or domino-like model. Then, where the high fire efficiency regions distributed, and how the relationship between fire efficiency and replication timing. How does ORM compare with other methods? What about the genetic functional annotation and epigenetic modification for initial zones? All these issues will be revealed one by one in the content of this Ph.D. study.

In the following part of my Ph.D. thesis, firstly, I will detail explain the Methods and Materials in Chapter 2. I will introduce 4 kinds of the basic data formats of the Bionano platform and how to calculate the fire efficiency based on ORM signals extracted from these 4 kinds of basic data and my developed packages. I will also introduce some technical problems of ORM technology, such as hot spots and the solution for the problem. In addition, we designed the experiments to observe the time-dependent movement of ongoing replication forks to test the reliability of ORM, prove almost all ORM signals that we detect are real signals reflecting ongoing replication regions, and use ORM signal density/fire efficiency to call and select initial zones. In Chapter 3, then I will present the optimal algorithm of initial zone calling, filtering, and testing results.

Besides the fire efficiency, the ORM signal intensity can also tell us something about the dynamic information of replication forks. I found the signal polarity which can provide the orientation of moving forks and observed the cell line-specific signal polarity in asynchronous data. These results will be shown in Chapter 4.

Last but not least, in Chapter 5 all analysis results around fire efficiency and the initial zones will be presented including the stochastic model simulation, epigenetic modification enrichment of IZ, the genetic function annotation of IZ, the comparison of different methods by ORC. Finally, I will give, in Chapter 6, a summary of current works and a perspective for new ORM methods and future studies.

CHAPTER 2

Material and Methods, and basic ORM signal analyses

2.1 Cell lines

This study used human HeLa S3 (RRID: CVCL_0058) and H9 cell lines (RRID: CVCL_9773). The wet lab experimental parts were performed by Kyle Klein at David M Gilbert's lab (Florida State University, USA), our collaborators in this study. Both lines are derived from female donors. HeLa S3 cells were grown in DMEM plus 10% Cosmic Calf Serum (GE Life Science SH30087) and Pen/Strep. H9 hESCs were grown in feeder free conditions on Geltrex matrix (Thermo Fisher A14133) coated dishes in StemPro (Thermo Fisher A100701) media according to manufacturer's specifications.

2.1.1 Cell synchronization

One of our major purposes is to detect DNA replication origins along the human genome, while the labeling regions in unsynchronized cells could be dominated by the trace of ongoing replication forks far from replication origins. Therefore, making sure that all cells start to replicate at the beginning of S phase can avoid such background noise resulted from separated replication forks, and increase the accuracy of origin signal calling in bulk data analysis.

In order to make sure all cells synchronized in G1/S border, we need to do cell arrest twice sequentially: firstly, arrested in the G2/M phase by nocodazole, and then at the G1/S phase border by aphidicolin. Because if only using aphidicolin alone, there are some cells, which have already entered into the S phase, will not be arrested at the G1/S border. This will introduce labeling noise to the final result.

Nocodazole is an antineoplastic agent, which can hinder the aggregation of microtubules. Microtubules are essential for cells to enter the mitotic phase so that cells stay in the G2/M phase of the cell cycle by this drug. In addition, aphidicolin is a tetracyclic diterpene antibiotic isolated from the fungus. It is a reversible inhibitor of DNA polymerase Alpha and Delta in eukaryotic cells, which can easily block the cell cycle at the G1/S border phase when using a very high concentration. After using nocodazole and aphidicolin in turn, cells were then spun down and washed 3 times with warm PBS (Phosphate-buffered saline) buffer to remove aphidicolin and make all cells enter into the S phase together. In our primary experiment, 87% of cells were synchronized in mitosis and, after release from aphidicolin, 94% incorporated the fluorescent label, with 100% of labeled cells showing an early-replication pattern of replication foci (Table 2.1, Figures S1.A-C), indicating a high degree of synchrony and <1% of contaminating asynchronous S phase cells.

2.1.2 Cell Labeling

We mainly report analyses concerning the characterization of labeling replication tracks and DNA fiber observed by the Bionano platform in 11 Hela S3 synchronized datasets and 2 Hela S3 asynchronous datasets, and 1 H9 asynchronous dataset. **Table 2.1** recapitulates the main information, including the cell lines, labeling properties, and detailed information about the alignment of DNA fibers. For observing the movement of labeling ORM signals at the different time points after S phase entrance, we wait for the indicated times (5, 10, 20, 30, 45, 60, or 90 minutes) to label the ongoing replication regions after cells enter into S phase. In each experiment (i.e. asynchronous HeLa or H9 cells, and synchronized HeLa cells at an indicated time point after released), cells were trypsinized and electroporated (Lonza, Nucleofection kit SE, HeLa S3 HV program) in the presence of 40 μM Aminoallyl-dUTP-ATTO-647N (Jena Bioscience NU-803-647N), the dUTP will mark replicating regions as dispersed red fluorescent signals.

2.2 Optical Replication Mapping

As mentioned in the introduction of Bionano high throughput mapping, there are 2 kinds of green mapping sites: NLRS (Nick, Label, Repair, and Stains) and DLS (Direct Label and Stain), both of them represent the specific motif sequence. The difference is that NLRS needs to cut and embeds the green fluorophore into the DNA molecule in NLRS motif sites (such as Nt.BspQI sites used in our study), then repairs at the fracture, which may cause the fragmentation of DNA molecules. But DLS is similar to the process of genetic modification, in which the enzyme DLE-1 (Bionano Genomics 80005) can mark corresponding DLE-1 motif sites without damaging DNA molecules. To some extent, DLS has better labeling performance than NLRS, with longer fibers and high mapped rates (Table 2.1).

When the corresponding motif sites of DNA molecules are marked with the green fluorescent group and the ongoing replication regions are also labeled by fluorescent dUTP, all isolated DNA samples will be loaded onto a Saphyr chip (Bionano Genomics 20319) and stretch to the linear structure on the corresponding Saphyr instrument. DNA fibers will be coaxed into parallel nanochannel arrays by electrophoresis, thereby extending the DNA to a uniform profile length, so that the distance along the molecule can be accurately measured. The molecules are imaged to collect YOYO-1 DNA signals in the blue channel, Nt.BspQI or DLE-1 motif sites in the green channel, and labeling replication regions *in vivo* in the red channel. They then pass through optical filters of different colors to collect corresponding red or green signals separately. Finally, the length of DNA fibers, the relative position of the red signals and the green signals on the fibers, and the signal intensity of the red and green signals will be recorded and saved in a BNX file. Through Bionano's matching algorithm, an XMP file will be further generated to record the positions of all fibers on the reference genome (human genome version hg19 is used in our study) and the genomic position information corresponding to each green signal.

Table 2.1 Basic information of all ORM datasets.

Sample Name	Cell Type	Time after synchronization for labeling	Synchronized %	% Alexa-dUTP Positive cells	% Early Foci	Green Mapping Labeling	Total Number of Fibers	Number of Maped Fibers	Mapped Percentage	% with ORM Signal
A.0	HeLa S3	0	87	94	100	NLRS	3992815	2915402	73	4
B.0	HeLa S3	0	87	69	100	DLS	3837247	3276738	85	8
C.0	HeLa S3	0	87	87	100	NLRS	5388744	4667642	87	7
C.5	HeLa S3	5	87	87	100	NLRS	1737248	1337663	77	9
C.10	HeLa S3	10	87	87	100	NLRS	1823407	1486208	82	8
D.0	HeLa S3	0	98	87	99	NLRS	1452292	911877	63	20
D.20	HeLa S3	20	98	95	98	NLRS	1521834	937385	62	25
D.30	HeLa S3	30	98	89	98	NLRS	1935712	832710	43	25
D.45	HeLa S3	45	98	88	97	NLRS	1599381	907477	57	32
D.60	HeLa S3	60	98	88	96	NLRS	1381105	993818	72	32
D.90	HeLa S3	90	98	91	97	NLRS	1415913	1022093	72	35
E.async	HeLa S3	Asynchronously labeled	**NA**	41	56	DLS	1347722	1190487	88	9
F.async	HeLa S3	Asynchronously labeled	**NA**	44	55	DLS	4798206	3759636	78	7

2.3 Data format of Bionano

There are 4 types of basic data format in Bionano output. They are .bnx, .rcmap, .qcmap and .xmp files. In the following paragraphs, I will introduce, respectively, a part of parameters within these four data formats involved in our data analysis process.

2.3.1 BNX

```
0	183	241500	4321.49	80.67	66	12967	1	-1	chips,SN_CDLERX6NPNRX7NW
U,Run_6c1a79ef-141a-4096-9d82-314634ab0357,0	1	1	4	4	405	813
4	399	1456
1	8431	15317	18719	20337	22378	23461	26511	38148	43239	44940	46152
48618	50418	55040	58304	62986	65026	67284	73302	75041	79238	83161	86625
90744	93541	98059	100275	103807	106615	108569	113238	115342	119415	123874	127147
128836	132386	136947	139303	143796	146698	148737	153243	156795	158939	163425	166261
169608	171133	173298	180030	182590	187506	189308	191170	194120	199178	205613	212731
215843	218206	221351	224684	229039	233971	238808	241500

2	372	30454	33754	53263	68284	73545	75769	80890	85650	87569	121950
132445	134124	161983	164214	183781	205368	208468	213711	227576	233488	239277	244203
252851	256800	264283	273708	277056	279563	285553	290158	297320	310045	311687	326948
336634	339543	344722	348611	352278	357012	358578	362169	364573	366917	368454	372675
376614	385439	387369	389689	395302	399662	405791	409123	416338	418047	423318	425846
429461	438289	442342	444530	464606	467240	472875

QX11	10.08348	6.33244	6.25445	6.27364	4.90548	6.82648	3.94645	20.34334	8.32105
7.80591	9.32729	11.34582	9.53998	7.06546	11.75305	11.87323	5.79447	6.69021
8.47028	12.94812	14.33600	8.09798	21.59392	5.61965	21.69473	21.09607
19.16816	9.50938	34.36756	11.66572	30.09792	41.02215
73.67329	17.74068	29.42975	39.03440	28.29641	28.13300
21.77220	30.53836	12.88604	5.35750	41.67753	17.60822	16.60717
7.78603	13.97649	11.65518	21.52665	15.09029	17.83440
16.97878	14.53676	12.57979	15.61269	15.79564	7.96846	28.43458
16.56446	4.34261	8.42204	14.80289	15.17917	12.42019	4.78720	6.89824

QX21	7.64803	9.44464	14.81698	8.44297	12.80191	12.53211	9.51094	6.65061
10.88102	9.18963	11.08927	10.47844	18.68086	5.27135	8.22451	9.80301
7.41648	15.81695	14.93037	13.41137	12.47227	9.26461	32.88548
19.00489	13.21570	10.60119	10.19845	14.87359	14.36357
9.82766	14.30509	20.68182	8.14954	17.95187	10.97401	16.51873
10.16308	33.78429	17.33000	5.89374	21.92048	8.46575	22.60254
18.22422	18.64907	11.70299	23.17835	9.13149	8.74133	14.90674
8.69611	9.19728	9.79332	15.49938	15.17399	12.41004	13.67709	17.42826
22.22139	9.82052	11.87555	22.77458	12.69888	8.88651	12.80701

QX12	583.12	366.20	361.69	362.80	283.68	394.77	228.22	1176.44	481.20	451.41	539.39
656.12	551.69	408.59	679.67	686.62	335.09	386.89	489.83	748.78	829.04	468.30	1248.76
324.98	1254.59	1219.97	1108.48	549.92	1987.45	674.62	1740.54	2372.28	4260.47	1025.93	1701.90
2257.33	1636.36	1626.91	1259.07	1766.01	745.19	309.82	2410.18	1018.27	960.38	450.26	808.25
674.01	1244.87	872.66	1031.35	981.87	840.65	727.48	902.87	913.45	460.81	1644.35	957.91
251.13	487.04	856.04	877.80	718.25	276.84	398.92

QX22	449.87	555.55	871.56	496.63	753.03	737.16	559.45	391.20	640.04	540.55	652.29
616.36	1098.84	310.07	483.78	576.63	436.25	930.38	878.23	788.88	733.64	544.96	1934.38
1117.90	777.37	623.58	599.89	874.89	844.89	578.08	841.45	1216.54	479.37	1055.96	645.51
971.66	597.81	1987.25	1019.38	346.68	1289.40	497.97	1329.52	1071.98	1096.97	688.39	1363.39
537.13	514.18	876.84	511.52	541.00	576.06	911.70	892.56	729.98	804.51	1025.16	1307.10
577.66	698.54	1339.64	746.97	522.72	753.33
```

Figure 2.1. Format diagram for BNX. Seven lines record 1 DNA fiber's information. The items in each line are described in detail in the main text. They are 0, 1, 2, QX11, QX21, QX12, QX22 in order. Please pay attention that there is no obvious line space between the line starts with 0 and the line starts with 1 at the given example.

.BNX file records the most primitive DNA-related information and does not involve any matching algorithms (Fig. 2.1). We only extract the key information for calculation. It will be introduced in detail below.

1. The line starts with 0 corresponds to the molecule backbone channel, and the second and third items of this line record the molecular ID and DNA fiber length, respectively. They are the only information we need to extract for each DNA fiber. In the above example, the molecular ID is 183, and the raw fiber length before recalibration is 241,500 bp.
2. The second and third lines start with 1 and 2, respectively, represent the 2 channels recording positions of green fluorescent group or positions of red labeling signals (florescent dUTP in our case). Users can choose different channels for saving green mapping signals and red labeling signals. So, in a different experiment, channel 1 and channel 2 could record different signals. All position information in these 2 channels will be recorded for downstream calculations.
3. QX11, QX21 record, respectively, the original signal intensity and SNR (signal-noise ratio) in channel 1. Similarly, the QX12, QX22 is the same information for channel 2, respectively. All intensity and signal noise ratio information in these 2 channels will be recorded for downstream calculations.

2.3.2 Rcmap and Qcmap

#h CMapId	ContigLength	NumSites	SiteID	LabelChannel	Position	StdDev	Coverage	Occurrence	GmeanSNR	lnSNRsd
#f int	float	int	int	int	float	float	float	float	float	float
1	1346021.3	305	1	1	7800.2	0.0	1.0	1.0	42.4000	0.0000
1	1346021.3	305	2	1	11447.6	0.0	1.0	1.0	15.6100	0.0000
1	1346021.3	305	3	1	17769.8	0.0	1.0	1.0	20.9300	0.0000
1	1346021.3	305	4	1	20969.2	0.0	1.0	1.0	32.8500	0.0000
1	1346021.3	305	5	1	23252.3	0.0	1.0	1.0	42.2100	0.0000
1	1346021.3	305	6	1	31189.0	0.0	1.0	1.0	24.7300	0.0000
1	1346021.3	305	7	1	33194.0	0.0	1.0	1.0	8.0200	0.0000
1	1346021.3	305	8	1	35419.5	0.0	1.0	1.0	32.3900	0.0000
1	1346021.3	305	9	1	39855.5	0.0	1.0	1.0	31.6900	0.0000
1	1346021.3	305	10	1	44453.6	0.0	1.0	1.0	30.8600	0.0000
1	1346021.3	305	11	1	47081.3	0.0	1.0	1.0	19.0500	0.0000
1	1346021.3	305	12	1	49820.5	0.0	1.0	1.0	29.1700	0.0000
1	1346021.3	305	13	1	51889.2	0.0	1.0	1.0	31.8600	0.0000
1	1346021.3	305	14	2	56803.8	0.0	1.0	1.0	23.8700	0.0000
1	1346021.3	305	15	1	61123.7	0.0	1.0	1.0	25.2700	0.0000
1	1346021.3	305	16	1	64062.7	0.0	1.0	1.0	35.7800	0.0000
1	1346021.3	305	17	1	67160.7	0.0	1.0	1.0	27.2400	0.0000
1	1346021.3	305	18	2	70297.8	0.0	1.0	1.0	8.5500	0.0000

Figure 2.2. Format diagram for .qcmap/.rcmap. The above plot just shows a small part of signal information for one DNA fiber. The first line is the column names that are described in detail in the main text.

Both .rcmap and .qcmap share the same data format and record all Nt.BspQI or DLE-1 motif sites related to the mapping (Fig. 2.2). The difference is that the .rcmap file records all motif sites in the reference genome, and the .qcmap file records green signals represent motif sites on DNA fibers. The columns in this file include the items below.

1. The CMapId is the same as the Molecular ID in BNX and the QryContig ID in .Xmap file.
2. The contiglenth is the DNA fiber length after recalibration. So, it may be slightly different from the DNA fiber length recorded in bnx file.

3. NumberSite records the signal number in this fiber. In the above example, there are 305 sites on this fiber with ID 1, including green mapping signals and red labeling signals. Each line here is a signal (1 or 2).
4. The SiteID is the ID for each signal it will be recorded sequentially based on the orientation recorded in Xmp.
5. The Labelchanel is the channel where the signal comes from. This is the way to distinguish whether it is a red labeling signal or a green mapping signal.
6. Position column records the genomic positions of signals.

All of these columns mentioned are extracted information that we need for downstream calculations.

2.3.3 Xmp

#h XmapEntryID	QryContigID	RefContigID	QryStartPos	QryEndPos	RefStartPos	RefEndPos	Orientation	Confidence	QryLen	RefLen	LabelChanne	Alignment
#f int	int	int	float	float	float	float	string	float	float	float	int	string
1	965410	1	9062.7	113649.8	14453.0	120827.0	+	17.49	376292.0	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(8,8)(11,9)(12,10)(13,11)(14,12)(15,12)(16,13)(17,14)(18,15)(19,16)(21,17)(22,18)(23,19)(24,19)
2	438774	1	611212.0	475612.6	14453.0	151828.0	-	22.65	618589.3	249250621.0	1	(1,67)(2,66)(3,65)(4,64)(5,63)(6,62)(8,61)(9,60)(10,59)(11,58)(12,57)(13,56)(14,55)(15,55)(16,54)(17,53)(18,52)(19,51)(21,50)(22,49
3	564432	1	264235.2	127081.2	14453.0	151828.0	-	24.92	271521.8	249250621.0	1	(1,36)(2,35)(3,34)(4,33)(5,32)(6,31)(7,30)(8,29)(9,28)(10,27)(11,26)(12,25)(13,24)(14,23)(15,23)(16,22)(17,20)(18,19)(20,18)(21,17)
4	834570	1	226526.6	89708.2	14453.0	151828.0	-	25.67	335212.6	249250621.0	1	(1,40)(2,39)(3,38)(4,37)(5,36)(6,35)(7,34)(8,33)(9,32)(10,31)(11,30)(12,29)(13,28)(14,27)(15,27)(16,26)(17,25)(18,24)(19,23)(20,22)
5	1454402	1	7806.3	152655.0	14453.0	157858.0	+	22.10	406087.6	249250621.0	1	(1,1)(2,2)(4,3)(5,4)(6,5)(7,6)(8,7)(10,8)(11,9)(12,10)(13,11)(15,12)(16,13)(17,15)(18,16)(19,17)(20,17)(21,18)(22,19)(23,20)(24,21)(
6	1732239	1	427471.4	286338.8	14453.0	157858.0	-	21.72	434899.0	249250621.0	1	(1,50)(2,49)(3,48)(4,47)(6,46)(7,45)(8,44)(9,43)(10,42)(12,41)(14,40)(15,40)(16,39)(17,38)(18,37)(19,36)(20,35)(21,34)(22,33)(23,32
7	1948414	1	8731.7	154597.3	14453.0	157858.0	+	21.84	238165.6	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(8,8)(9,9)(10,10)(11,11)(12,12)(13,13)(14,14)(15,14)(16,15)(17,16)(18,17)(20,18)(21,19)(22,20)(23,
8	1990219	1	7124.3	151836.2	14453.0	157858.0	+	25.69	200117.0	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(7,8)(8,9)(9,10)(10,11)(11,12)(12,13)(13,14)(14,15)(15,15)(16,16)(17,17)(18,18)(19,20)(20,21)(21,22)(22,23)(
9	57979	1	257078.0	111907.8	14453.0	160706.0	-	24.05	263601.1	249250621.0	1	(1,40)(2,39)(3,38)(5,37)(6,36)(7,35)(8,34)(9,33)(10,32)(11,31)(12,30)(13,29)(14,28)(15,28)(16,27)(17,25)(18,24)(19,23)(20,23)(21,22
10	1710934	1	559935.8	414590.4	14453.0	160706.0	-	27.91	567098.3	249250621.0	1	(1,58)(2,57)(3,56)(4,55)(5,54)(6,53)(7,52)(8,51)(9,49)(10,48)(11,47)(12,46)(13,45)(14,44)(15,44)(16,43)(17,41)(18,40)(19,39)(21,38)
11	1153455	1	9698.7	156528.6	14453.0	161634.0	+	25.63	159473.2	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(9,8)(11,9)(12,10)(13,11)(14,12)(15,12)(16,13)(17,14)(18,15)(19,16)(20,17)(21,18)(22,19)(23,20)(24
12	1665078	1	9581.4	156978.2	14453.0	161634.0	+	21.24	158559.8	249250621.0	1	(1,1)(2,2)(4,4)(5,5)(6,6)(7,7)(8,8)(9,9)(10,10)(11,11)(12,12)(13,13)(14,14)(15,14)(16,15)(17,16)(18,17)(19,18)(20,18)(21,19)(22,20)(
13	1436661	1	8536.5	161256.7	14453.0	166654.0	+	23.53	164040.1	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(8,8)(10,9)(11,10)(12,11)(13,12)(14,13)(15,13)(16,14)(17,15)(18,16)(19,17)(21,18)(23,19)(24,20)(25
14	1633715	1	9818.3	164443.7	14453.0	166654.0	+	22.13	224425.4	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(8,8)(9,9)(10,10)(11,11)(12,12)(13,13)(15,14)(17,15)(18,16)(20,17)(21,18)(22,19)(23,20)(24,21)(25,
15	2145171	1	152943.0	1982.7	14453.0	166654.0	-	20.30	161290.9	249250621.0	1	(1,28)(2,27)(3,26)(4,25)(5,24)(6,23)(8,22)(9,21)(10,20)(11,18)(12,17)(13,16)(14,15)(16,14)(17,13)(18,12)(19,9)(21,8)(22,7)(23,6)(25,
16	535401	1	8715.6	168649.1	14453.0	174178.0	+	22.01	223853.3	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(8,8)(9,9)(11,11)(12,12)(14,13)(15,13)(16,14)(17,15)(18,16)(19,17)(20,18)(21,19)(22,20)(23,21)(25,
17	806268	1	10441.6	170074.0	14453.0	174178.0	+	26.13	269848.9	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(8,8)(9,9)(10,10)(12,11)(13,12)(14,13)(15,13)(16,14)(17,15)(18,16)(19,17)(20,18)(21,19)(22,20)(24,
18	1141242	1	8786.0	166653.0	14453.0	174178.0	+	20.67	322527.3	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(9,8)(10,9)(11,10)(12,11)(13,12)(14,13)(15,13)(16,14)(17,15)(19,16)(20,16)(21,17)(22,18)(23,19)(24
19	1555129	1	371630.4	211939.7	14453.0	174178.0	-	25.95	377969.5	249250621.0	1	(1,43)(2,42)(3,41)(5,40)(6,39)(7,38)(8,37)(9,36)(10,35)(11,34)(13,33)(14,32)(15,32)(17,31)(18,30)(19,29)(21,28)(22,27)(23,26)(24,25
20	1661848	1	378544.3	218344.6	14453.0	174178.0	-	28.93	386559.1	249250621.0	1	(1,49)(2,48)(3,47)(4,46)(5,45)(6,44)(7,43)(8,42)(9,41)(10,40)(11,39)(12,38)(14,37)(15,37)(16,36)(17,35)(18,33)(19,32)(20,31)(21,30)
21	144105	1	9097.7	170208.2	14453.0	176619.0	+	22.81	217714.9	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(8,8)(9,9)(10,10)(11,11)(12,12)(13,13)(14,14)(15,14)(16,15)(17,16)(18,17)(19,18)(21,19)(22,21)(23,
22	216007	1	8901.8	170270.2	14453.0	176619.0	+	19.93	176360.1	249250621.0	1	(1,1)(2,2)(3,3)(5,5)(6,6)(7,7)(9,8)(10,10)(11,11)(12,12)(13,13)(14,14)(15,14)(18,15)(21,16)(22,17)(23,18)(24,19)(25,20)(26,21)(28,23
23	239861	1	6714.0	167170.7	14453.0	176619.0	+	18.78	194049.6	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(8,8)(9,9)(10,10)(11,11)(12,12)(17,16)(18,17)(21,18)(22,19)(23,20)(24,21)(25,22)(26,23)(28,25)(29,
24	377518	1	286372.3	125550.9	14453.0	176619.0	-	27.62	292721.0	249250621.0	1	(1,39)(2,38)(3,37)(5,36)(6,35)(7,34)(8,33)(9,32)(10,31)(11,30)(12,29)(13,28)(14,27)(15,27)(16,26)(17,25)(18,24)(19,23)(21,22)(22,21
25	395568	1	529159.4	367145.3	14453.0	176619.0	-	24.39	536726.8	249250621.0	1	(1,53)(2,52)(4,51)(5,50)(6,49)(7,48)(8,47)(9,46)(10,45)(11,44)(12,43)(13,42)(14,41)(15,41)(16,40)(17,39)(18,38)(20,36)(21,35)(22,34
26	423538	1	7727.1	170794.6	14453.0	176619.0	+	28.91	274108.1	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(6,5)(7,6)(8,7)(9,8)(10,9)(11,10)(12,11)(13,12)(14,13)(15,13)(16,14)(17,15)(18,16)(19,17)(21,18)(22,19)(23,20)(24
27	449862	1	462941.3	301043.9	14453.0	176619.0	-	28.44	470123.0	249250621.0	1	(1,54)(2,53)(3,52)(4,51)(5,50)(6,49)(7,48)(8,47)(9,46)(10,45)(11,44)(12,43)(13,42)(14,41)(15,41)(16,40)(17,38)(18,37)(19,36)(20,36)
28	480799	1	341196.1	176955.1	14453.0	176619.0	-	28.63	354012.7	249250621.0	1	(1,45)(2,44)(4,43)(5,42)(6,41)(7,40)(8,39)(9,38)(10,37)(11,36)(12,35)(13,34)(14,33)(15,33)(16,32)(17,31)(18,30)(19,29)(21,28)(22,27
29	603424	1	277963.0	115421.9	14453.0	176619.0	-	27.96	288738.6	249250621.0	1	(1,40)(2,39)(3,38)(4,37)(5,36)(6,35)(7,34)(8,33)(9,32)(10,31)(12,30)(14,29)(15,29)(16,28)(17,27)(18,26)(20,25)(21,24)(22,23)(23,22)
30	644372	1	10882.1	171560.3	14453.0	176619.0	+	26.60	304391.9	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(8,8)(9,9)(10,10)(11,11)(12,12)(13,13)(14,14)(15,14)(16,15)(18,16)(20,17)(21,18)(22,19)(23,20)(24,
31	705649	1	9136.9	172711.4	14453.0	176619.0	+	28.90	386453.9	249250621.0	1	(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)(7,7)(8,8)(9,9)(10,10)(12,11)(13,12)(14,13)(15,13)(16,14)(17,15)(18,16)(19,17)(20,17)(21,18)(22,19)(23,

Figure 2.3. Format diagram for xmap. Here gives an example of 30 records of mapping DNA fibers. The first line column names correspond to the key information that is described in detail in the main text.

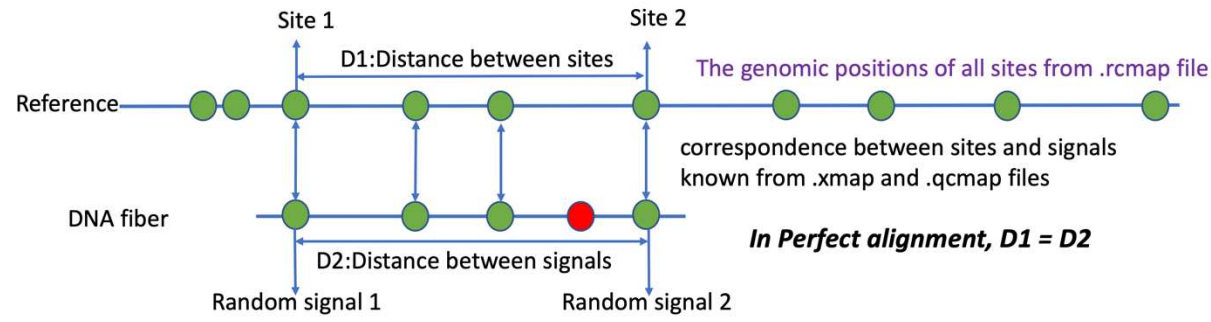
.Xmp file records all mapping results by Bionano algorithms (Fig. 2.3). Among the many complicated information encoding in the file, our analysis only extracts a few columns of key information for calculation including the items below.

1. QryContigID: The same molecular ID as in .BNX
2. RefContigID: The Chromosome, same as in .qcmap and .rcmap
3. QryStartPos: The coordinates of the first aligned labels on the query map, which corresponds to the relative position of the first green signal on DNA fiber,
4. QryEndPos: The coordinates of the last aligned label on the reference, which corresponds to the genomic position of the last green signal on DNA fiber.
5. Orientation: Orientation (+/-) of this DNA fiber in the order of first to last green signal is from left to right (+) or right to left (-) depending on which strand it mapped to.

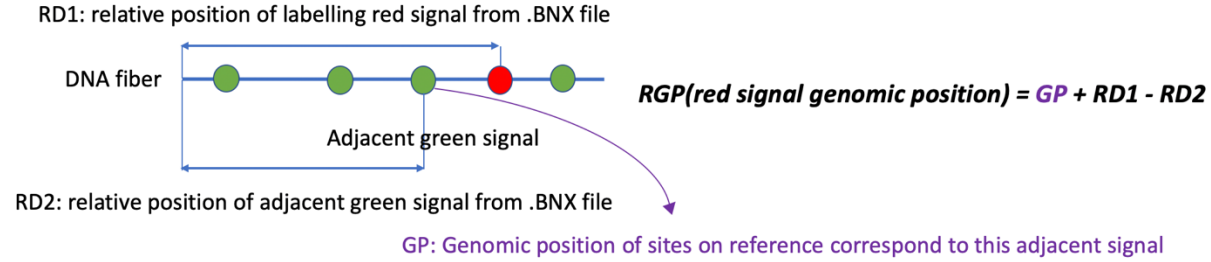
6. QryLen: The length of DNA molecule after recalibration. After the correction, there may be a slight discrepancy with the length of the corresponding DNA molecule in the BNX file.
7. Alignment: The motif sites pairing between reference sites and green signals on fibers. In the form of brackets, combine the motif site ID in r.cmap on the reference sequence and the ordinal number of the green fluorescent signal in q.cmap on the mapping DNA fiber. For example, the pair (32,14), firstly based on the Chromosome number of Xmap record, we can find all motif sites on the corresponding chromosome of .rcmap, and based on the molecular ID, we can find the corresponding record in BNX to get all relative positions of green mapping signals and red labeling signals, respectively, in BNX. The 32 is the motif site with the site ID 32 in the same chromosome of r.cmap. And 14 is the 14th green mapping signal. Such a one-to-one relationship represents the 14th green signal on DNA fiber mapped to the reference position with green signal has ID 32 of the corresponding chromosome.

2.4 The calculation of genomic positions for the red signals

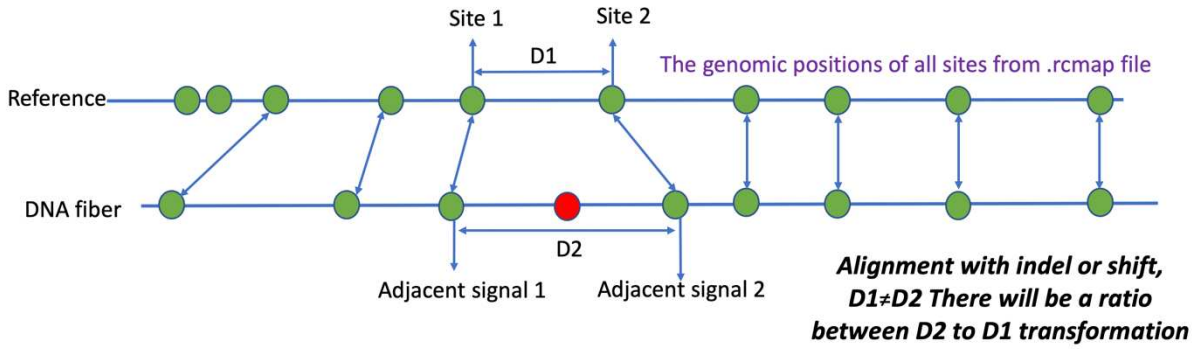
A1



A2



B1



B2

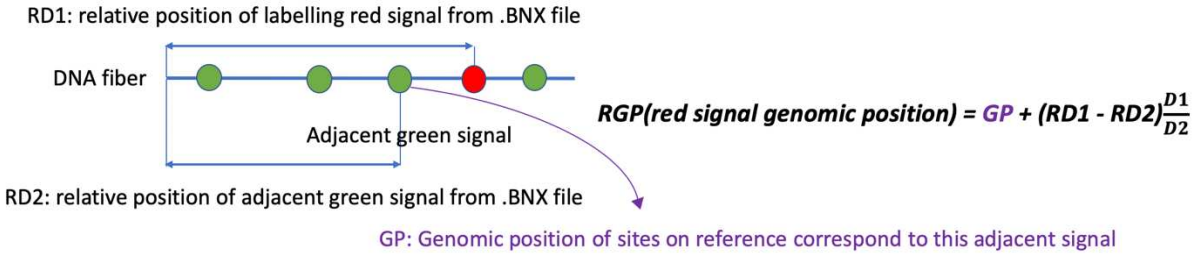


Figure 2.4. Two situations for signal mapping and their genomic position calculation. The blue lines represent reference DNA sequence or DNA molecule fiber; all solid green circles mean the green signals for mapping and red signals represent the labeling signals. The double arrow between two green circles means the corresponding relationship between them, and the double arrow between two arrows means the distance between them. All the sentences marked by purple indicate the information about the genomic position of sites on reference and the data source is .rcmap file. A1, A2 show an example where the signals on the reference genome and DNA fibers are perfectly aligned; and B1, B2 show an example with imperfect alignment.

As shown in Fig. 2.4, the genomic positions of ORM (red) signals need to be calculated based on the adjacent mapping signals. There are two common situations, which are detailed described below.

A1: For an ideal case in which DNA molecule is perfectly mapped to reference, all green signals in DNA fiber are fully fit the green site positions on the reference. The distance between 2 adjacent green signals is equal to the distance between 2 corresponding green sites on reference (D1 = D2). All genomic positions of mapped sites could be found in .rcmap file.

A2: Any red signal genomic position in perfect alignment case can be calculated by adjacent green signals. Firstly, the genomic position of the closet green signal is the same as the genomic position of corresponding site on the reference (GP marked by purple color). Since we also know the relative position of red signal (RD1) and closet green signal (RD2) from BNX file, the red signal genomic position (RGP) should be the genomic position of adjacent green signal (GP) add the difference of relative position of red signal as calculated in formula (2.1).

$$RGP = GP + RD1 - RD2 \quad (2.1)$$

B1: For the most real cases in which alignment shows a shift or small insertion or deletion (indel), compared with the distance between green signals on the fiber, the distance between corresponding

sites on reference may larger or smaller (therefore, D1≠D2). All genomic positions of sites on the reference are available from .rcmap file. But their genomic positions may be not equal to those of corresponding green signals.

B2: For each red signal, if indel or shift exists between 2 closet green signals to the corresponding sites on the reference. We can fix the difference of relative position by multiplying a ratio based on the distance between 2 nearby green signals to distance between 2 corresponding sites on the reference (D1/D2 in figure 2.4 B1). In this way, we can proximity restore the accurate genomic positions of red signals. So, the red signal's genomic position (RGP) is equal to GP + (RD1-RD2) x D1/D2 as indicated in formular (2.2).

$$RGP = GP + (RD1-RD2) \times D1/D2 \qquad (2.2)$$

2.5 Data integration by jar packages and output format

Based on the formula (2.2) in section 2.4, I developed a series of jar packages (https://github.com/CL-CHEN-Lab/ORM) to perform analyses from the 4 types of raw data (i.e. .bnx, .xmap, .qcmap and .rcmap). A detailed description of each jar package can be found in the supplementary material section (Table S1). In the following section. I will focus on the two most commonly used basic jar packages and their output format.

2.5.1 AllRawDataRefining.jar and its output format

Because the Bionano technology is originally developed for *de novo* assembly, here we applied this technology for the detection of replication origins in our ORM approach. AllRawDataRefining.jar is, therefore, developed to extract and reintegrate the original data and calculate the genomic positions of ORM red labeling signals. from 4 input data, .bnx (raw labeling red signals data), .xmap (mapping data), .qcmap (base calibration data) , .rcmap (reference data) of a given sample. The jar package should be run under the terminal. It will integrate all information from these 4 input files, such as filtering unmapped fibers or fibers without any labeling signals and calculating the precise genomic positions of ORM labeling signals. Normally, there are 2 channels, one stores the green signals for mapping fibers to reference, another is used to store the labeling red signals (Fig. 2.1 and 2.4). Different experiments may have different choices for labeling. So, the user needs to set which channel used for ORM signal labeling and which channel used for mapping by options -S and -M, respectively, and whether they want to restore the information of the mapping green signals (genomic position, signal intensity, and signal noise ratio) for further analysis in .txt output files by option -WGI. The detailed function and command-line example can be checked by manual of jar packages on the GitHub page: https://github.com/CL-CHEN-Lab/ORM/blob/master/User.Manual.docx .

In order to record the summary of all calculation results, the jar package will generate a .txt file like Figure 2.5. In this jar package script, the users can choose the channels that they are interested in based on the research purposes. So, the .txt output format and content could be different for

different parameter settings in a script running. Next, I will introduce the parameters that affect the output results and the differences in the output results

2.5.1.1 The output only contains the information of red signals

Generally, what we are interested in is only the genomic positions of the ORM labeling signals. When we set the parameter option **-WGI** as N instead of Y, the .txt output will be like below (Fig. 2.5).

```
1905.FC0_26     Chr22   -
17539192        18476114        18617660
0       209.69  0
0       11.12   0

1905.FC0_61     Chr8    -
76441528        76733539        77156379
0       74.02   0
0       3.97    0

1905.FC0_87     Chr7    -
134773559       135281253       135484282
0       126.49  0
0       6.96    0
```

Figure 2.5. An eample of three fiber records in .txt files when -WGI set as "N".

In this output file, each fiber record is organized into 4 lines. The first item in the 1st line is a record ID organized by sample name and molecular ID file separated by "_", the second item is chromosome, and the third one indicates the orientation that the fiber mapped to the reference genome: from left to right (+) or right to left (-). Both molecular ID and chromosome come from .BNX file and the sign of alignment comes from .xmap file.

The second line provides the calculated genomic positions of all labeling red signals and calculated genomic coordinates of 2 ends of the corresponding fiber. As formula (2.2) shows, the coordinates of the red signal need to take the closest green signal as a reference to calculate. The same goes for both ends of the DNA fiber. 2 fiber ends can take the first and last green signals on fiber as closet green signals, respectively, using their genomic position (GP) from .rcmap and relative position from .bnx (RD1). Then the formula 2.2 still needs RD2 to calculate the genomic position of fiber's ends. The two ends of fiber will take 0 and the recalibrated DNA fiber's length from .qcmap as relative position on fiber (RD2), and substitute them into formula (2) in section 2.4 to calculate.

The 3rd and 4th lines are corresponding SNR (signal-noise ratio) and signal intensity, respectively, from the .bnx file. Since the first and last items in the second line are the genomic positions of fiber ends instead of labeling signals, their corresponding values in the 3rd and 4th lines are 0.

2.5.1.2 The output contains information for both green and red signals

Sometimes, we also want the information of green signals. For this, we only need to set the **-WGI** parameter as "Y", the output files will be like those shown in Fig. 2.6. For example, we observed that there are some false positive red signals that appear at specific genomic regions, named hot spots. These hot spot red signals are associated with the green mapping signals. The out .txt file with the information of green signals is useful to analyze these hot spot signals, e.g. to prove their existence, their relation with the green signals, and the filtering of the hot spot signals. The related analysis will be further described in section 2.6.

```
0min_376        Chr9    +
100223873       100259542       100263454       100270363       100285221       100297683       100310230       100618780
0       668.59  248.12  188.02  250.42  349.9   222.73  0
0       32.46   12.05   9.13    12.16   16.99   10.81   0
100223873       100245575       100245575       100248085       100261848       100277539       100281313       100286543
100298370       100301804       100311310       100319080       100324964       100331514       100335513       100347773
100363473       100367326       100376943       100379508       100396285       100399326       100402951       100407313
100413043       100426758       100441863       100445905       100450299       100461829       100479736       100481824
100487639       100495873       100499218       100503372       100515409       100520787       100522509       100528883
100542311       100543716       100547006       100551380       100553702       100560119       100563859       100578193
100580052       100582844       100614035       100616254       100618780
0       1649.36 635.0   752.48  3992.78 326.86  1832.08 825.43  5006.65 1569.2  612.06  1356.83 3776.18 1316.93 1210.36 785.27  2211.25
        1340.43 1957.16 1661.26 1357.82 342.16  580.65  396.95  385.69  664.49  828.86  357.08  296.53  1146.13 1229.07 1160.96 293.1
        584.64  1421.92 3575.45 287.58  614.57  2165.75 2809.55 2080.41 1482.76 1024.69 417.57  280.05  742.7   487.33  2220.04 439.5
        2022.38 211.49  234.22  2104.31 4412.26 0
0       85.32   32.85   38.92   206.53  16.91   94.77   42.7    258.98  81.17   31.66   70.18   195.33  68.12   62.61   40.62   114.38
        69.34   101.24  85.93   70.24   17.7    30.03   20.53   19.95   34.37   42.87   18.47   15.34   59.28   63.58   60.05   15.16
        30.24   73.55   184.94  14.88   31.79   112.03  145.33  107.61  76.7    53.0    21.6    14.49   38.42   25.21   114.83  22.73
        104.61  10.94   12.12   108.85  228.23  0
```

Figure 2.6. Example of a fiber record in .txt file when -WGI set as "Y". Due to the limit of display item length, it crowded into 19 lines, but in fact, there are only 7 lines. Each fiber record is organized by 7 lines. The first 4 lines are the same as in .txt output file with "-WGI" parameter set as "N". And the 5th to 7th lines are the positions for green mapping signals, SNR, and intensity, respectively.

2.5.1.3 The output contains fiber without any red signal

There are often fibers that do not carry any red signal. Even when doing hot spot filtering, the intention is for comparing the intensities of red signals and green signals, and their distribution difference. So, for the sake of simplicity of the results, these fibers will be automatically filtered with the default parameters. But if the user needs to evaluate the labeling rate or normalize the labeling signal based on the number of mapped fibers, we need also these fibers without any ORM signal. This time, we need to set the -WNS set as "Y". The output will contain more records like below, here the example is with the -WGI as "Y", too (Fig. 2.7).

```
0min_4  Chr8    +
87234086        88223911
0       0
0       0
87234086        88223911
0       0
0       0
```

Figure 2.7. One record of fiber without any labeling signals in .txt files when -WNS as "Y" and -WGI set as "Y". Such records are for users who need the information of fibers without ORM signal. To safe storage space, no matter for green signals or red signals, the output file only records two ends of fibers even if the green signals exist on this fiber. And there is no intensity and SNR information in output files.

2.5.2 GenerateGTF_ByAllDataRefining_Reformat.jar and its output format

In order to better observe the signal distribution of labeling signals, we generate a gtf-like file (Fig. 2.8) by jar package：GenerateGTF_ByAllDataRefining_Reformat.jar.

1	processed_transcript	transcript	4056	247531
1	processed_transcript	exon	4056	4056 .
1	processed_transcript	exon	24488	25488 .
1	processed_transcript	exon	28048	29048 .
1	processed_transcript	exon	33970	34970 .
1	processed_transcript	exon	38765	39765 .
1	processed_transcript	exon	43815	44815 .
1	processed_transcript	exon	46397	47397 .
1	processed_transcript	exon	51760	52760 .
1	processed_transcript	exon	53122	54122 .
1	processed_transcript	exon	59953	60953 .
1	processed_transcript	exon	87930	88930 .
1	processed_transcript	exon	106318	107318 .
1	processed_transcript	exon	127811	128811 .
1	processed_transcript	exon	129256	130256 .
1	processed_transcript	exon	247531	247531 .

Figure 2.8. GTF-like file for all labeling signals on fibers. Within this gtf-like file, transcript means DNA fiber and exon means the labeling red signal and two ends of DNA fiber. For data visualization, the first and last exon in one transcript represents the 2 ends of fibers, in the above example, we can see their start and end positions are the same (4056 and 247531) in order to show the total length of DNA fiber. For each exon that represents red labeling signals, we extend 500bp to 2 sides. The 1kb exon region will show as a little bar in IGV to represent the relative position in each DNA fiber.

All fibers will be recorded in the form of a transcript, the two ends of fibers, and all ORM signals will be recorded as exons. Besides two ends of fibers, all red signals will become 1 kb extended areas centered on the genomic position of each ORM red signal. In this way, we can put such a gtf-like file under IGV (or other genomic browsers) for visualization (Fig. 2.9).

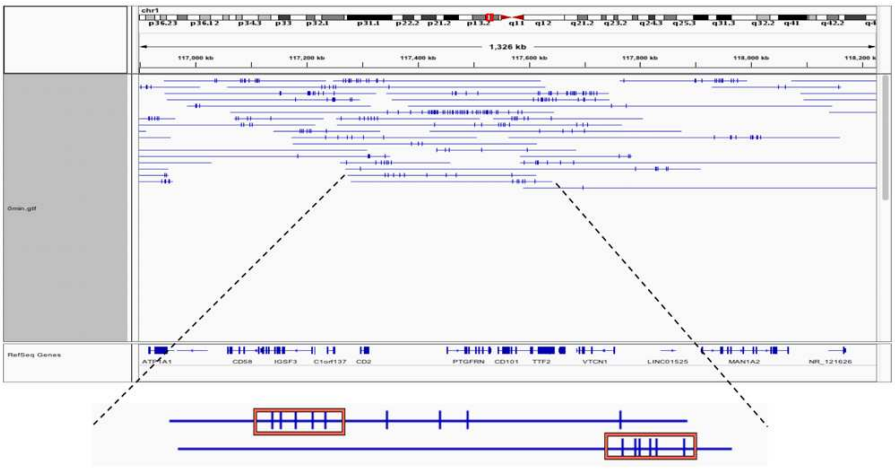


Figure 2.9 Visualization of ORM signals on the fibers with IGV by using the gtf-like file. Each line is a DNA fiber, and each bar is a labeling red signal. Close adjacent signals with distances smaller than cutoff values defined by users will be clustered as an ORM segment in the following analysis (see section 2.7), like the signals enclosed by red frames in the zoom-in plot.

2.6 Hot spots filtering

2.6.1 Hot spots

When we analyzed the red signal distribution of raw data visualization of 0' samples along the genome (Fig. 2.9), we observed a strange signal enrichment in some specific positions (Fig. 2.10). We called such kind of abnormal enrichment ORM signals hot spots.

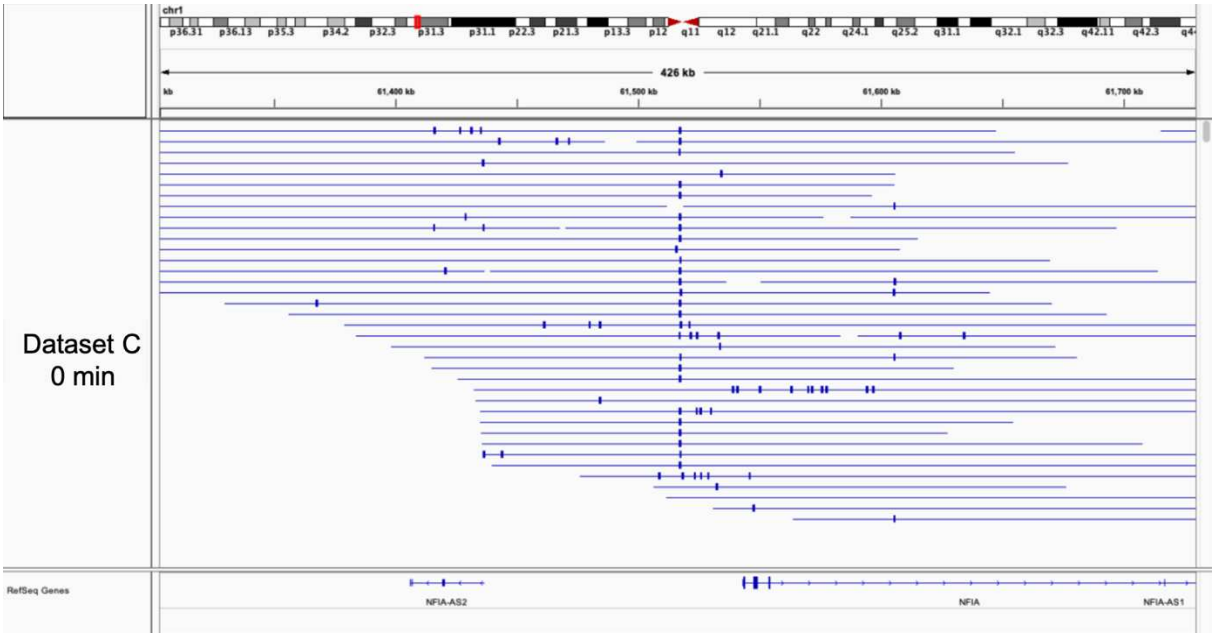


Figure 2.10. An example of abnormal ORM signal enrichment around specific sites of a 0 min ORM sample. The fibers and ORM signals are visualized on IGV as in Fig. 2.9.

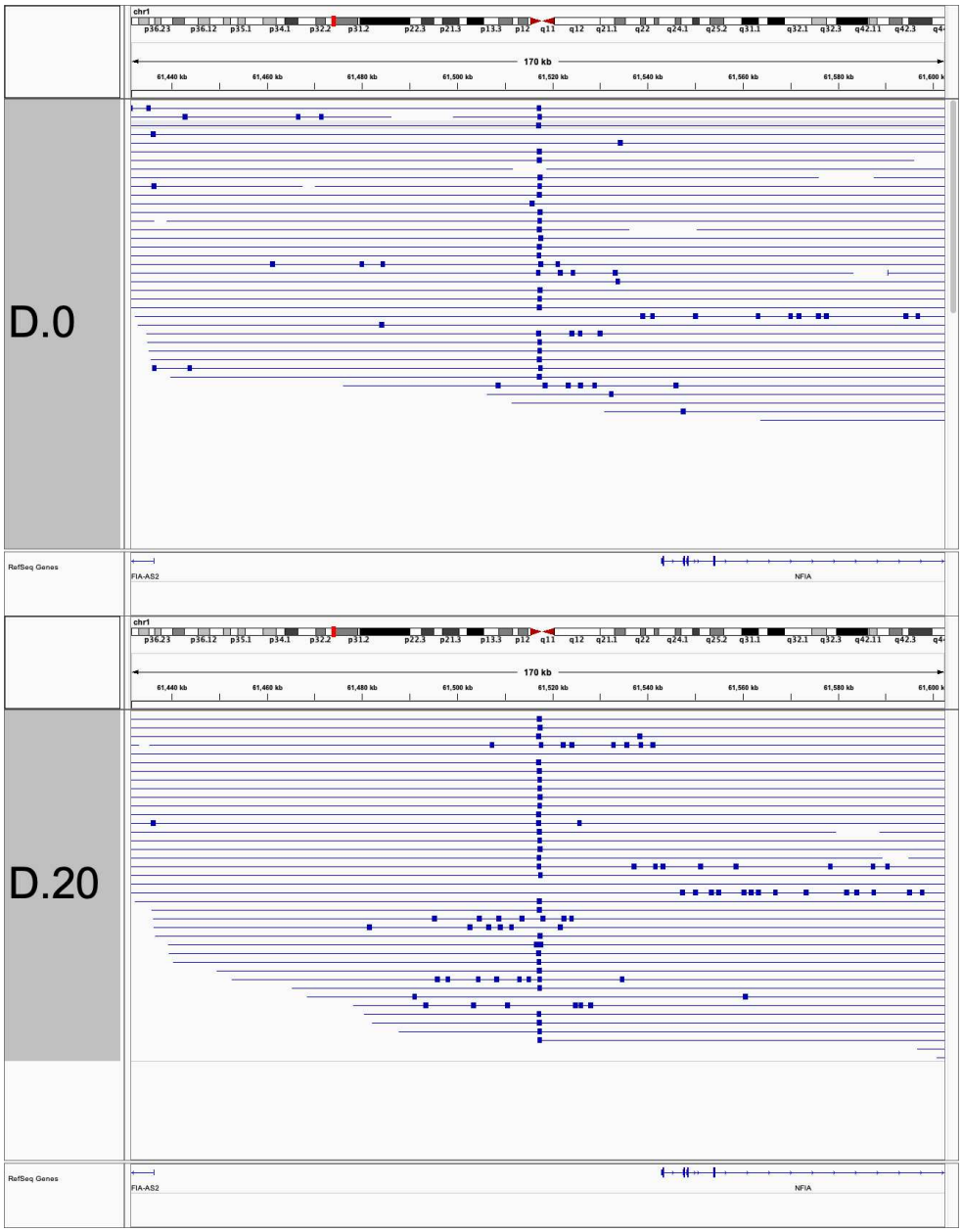
If these ORM signals located at the hot spots observed at 0' samples are real replication signals associated with replication regions, they should correspond to site-specific replication origins. Therefore, the positions must change and disappear with the movement of the replication process within the ORM data of different time points after the S phase entrance.

However, the accumulation of ORM signals at the same positions occurred in all 0 min, 5 min,10 min samples of dataset C and 0~90 min samples of dataset D (C.0~C.10, D.0~ D.90 in Table 2.1) (Fig. 2.11). Therefore, it strongly suggests that the hot spots should not be real biological signals related to replication initiation.

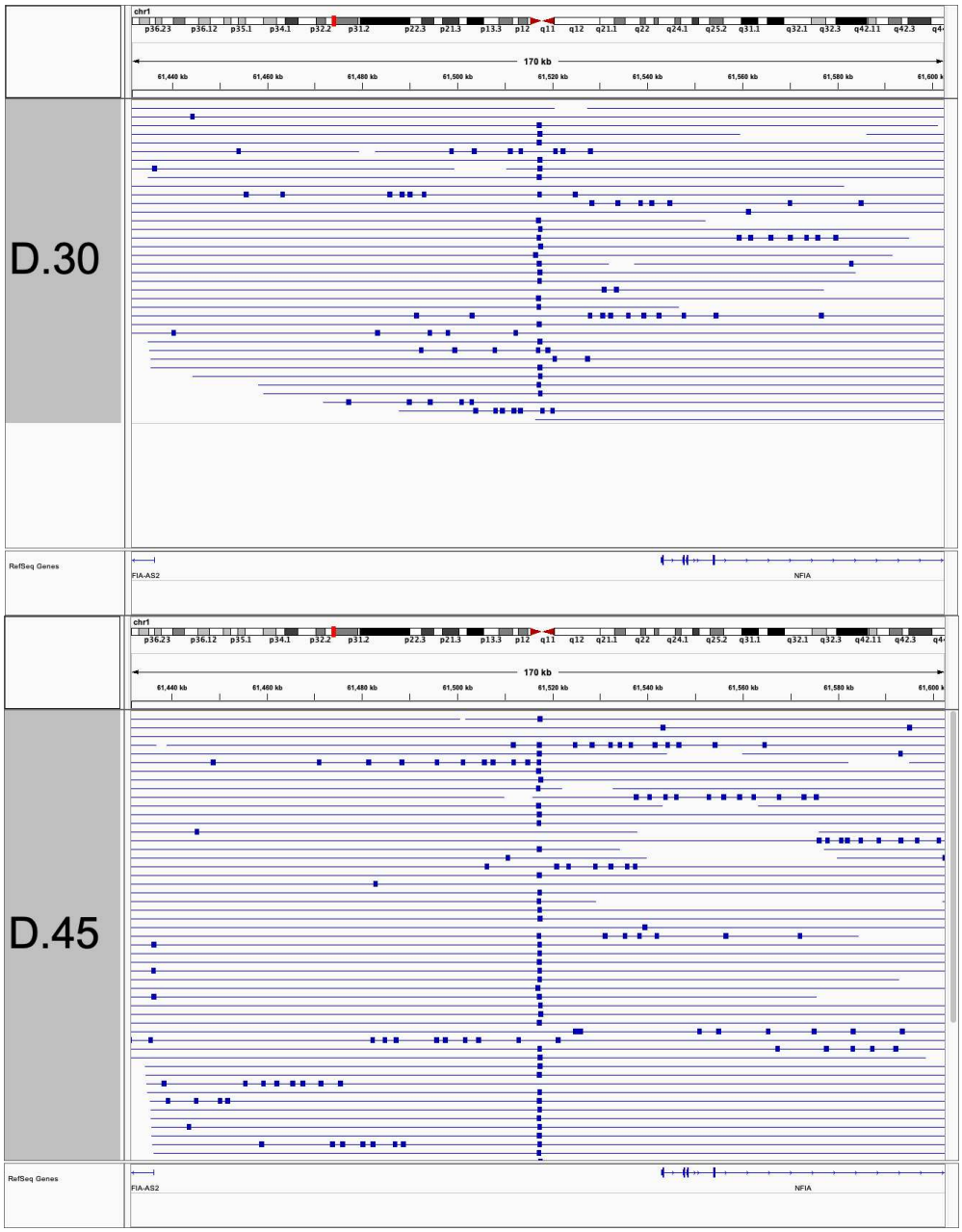
After a detailed investigation, we found that most of the abnormal signal accumulation only occurred around positions mapped by NLRS methods. Both datasets C and D were used NLRS to align DNA fibers to the reference genome (NLRS in labeling column of Table 2.1), and the datasets used DLS (dataset B) don't have such a strange signal enrichment phenomenon (Fig. 2.12). We thus speculated that these hot spots might result from too strong mapping green signals passing red optical filter in NLRS labeling. To confirm the hot spot enrichment is just caused by NLRS site instead of DLS site. We decide to choose an observation range containing both DLS and NLRS sites to see if we can only observe the hotspot distribution in the dataset using NLRS. Because the

position of all DLS and NLRS sites are fixed and recorded in r.cmap file of the different dataset. As shown in Fig 2.12, within this region from 61510828 to 61518030 bp on chromosome 1, it contains one NLRS site at 61517359 bp (associated with dataset D.0) and two DLE1 sites at 61510828 bp, 61518030 bp (associated with dataset B.0), respectively. As we expected, only dataset D.0 containing the enriched hot spots, which is associated with the NLRS site. There is no similar phenomenon that occurred in DLE1 sites of the B.0 sample (Fig. 2.12).

A: Beginning of S phase



B: Mid S phase progression



C: Later S phase progression

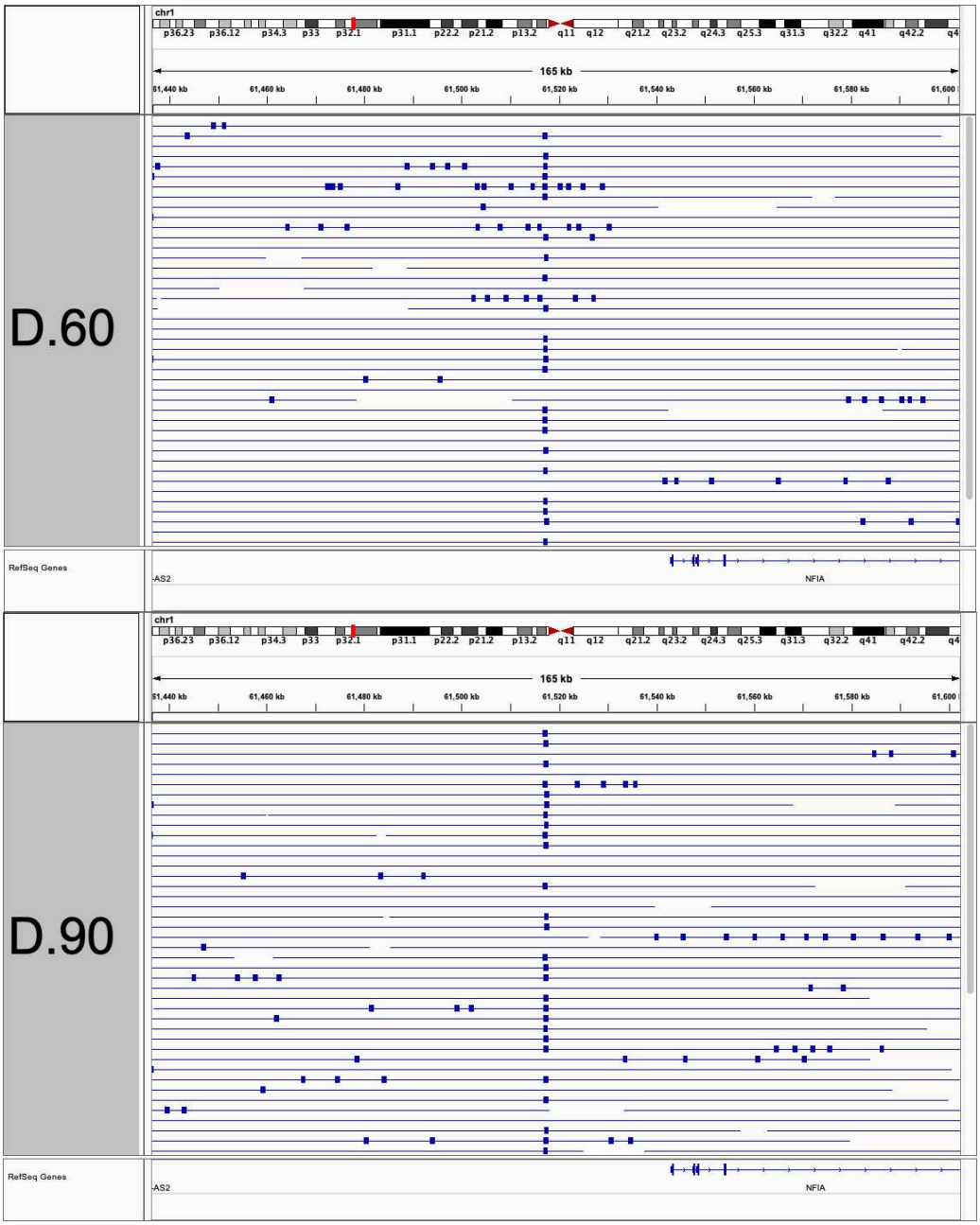


Figure 2.11. Abnormal ORM signal enrichment around specific sites at the beginning (A), Mid S phase progression (B), and later S phase progression (C). The specific time points after entry of the S phase are 0 min, 20 min, 30 min, 45 min, 60 min, and 90 min as indicated on the figures. The fibers and ORM signals are visualized on IGV as in Fig. 2.9.

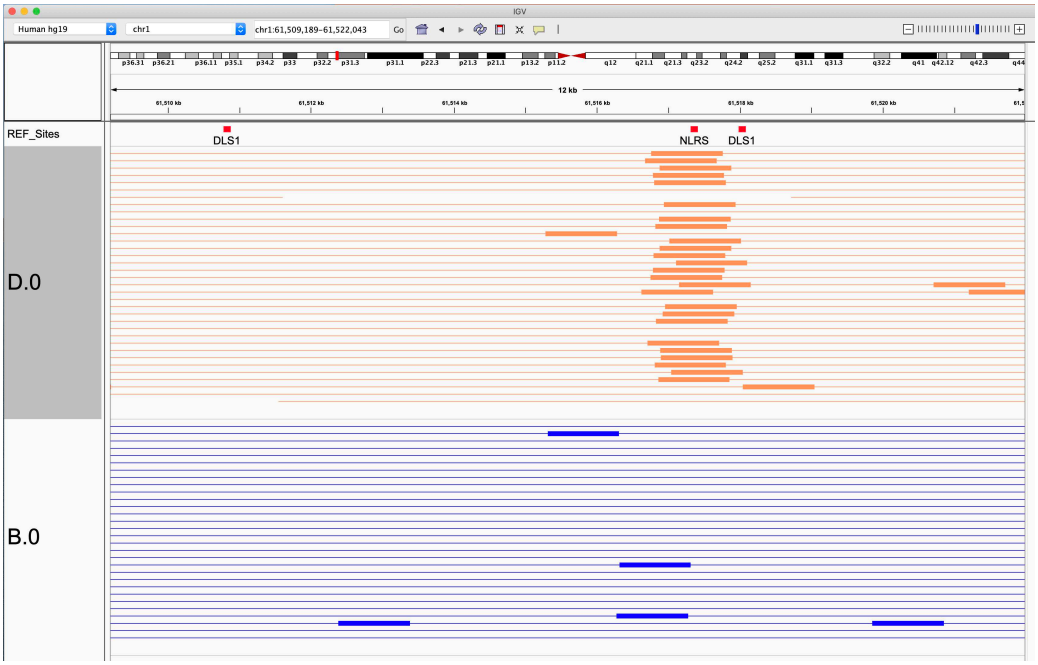


Figure 2.12. The zoom in abnormal signal enrichment of D.0 sample (using NLRS site for mapping) compared with B.0 sample (using DLS1 site for mapping). The first line red squares show the positions of 3 motif sites: 1 NLRS site and 2 DLS1 sites. The second and third lines show the ORM signal distribution of D.0 (orange) and B.0 (blue). The short framework on the solid lines represents zoom-in ORM signals on DNA fibers.

Then, we look through the entire genome of samples using the DLS method (Table 2.1) and didn't find the hot spot enrichment phenomenon as that occurred in datasets using the NLRS method. To further confirm that such hot spot signals might false-positive signals, I took the hot spot enriched site center ±300 bp regions as hot spot regions and detected the signal intensity distribution between signals inside and outside hot spot regions (Fig. 2.13 and 2.14).

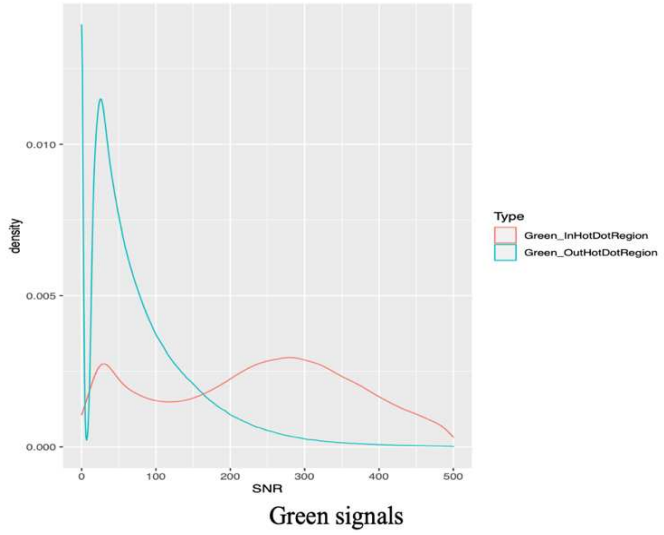


Figure 2.13. The density curve of SNR (signal-noise ratio) of green mapping signals. The red line is the distribution for green signals within the hot spot regions, and the blue one is for the green signals outside the hot spot regions.

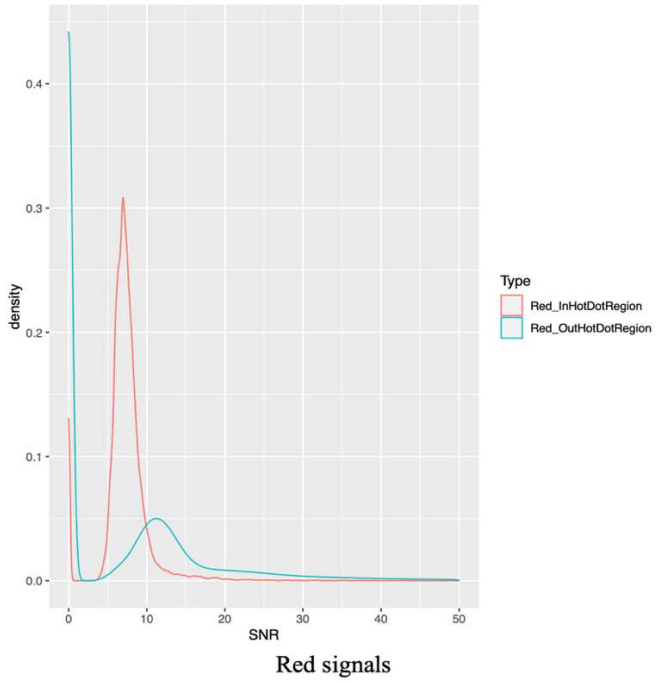


Figure 2.14. The density curve of SNR (signal-noise ratio) for red labeling signals. The red line is the distribution for red signals within the hot spot regions, and the blue one is for the red signals outsides the hot spot regions.

Based on figure 2.13 and 2.14, the green signal intensity in hotspot regions is stronger than the signals outside. Furthermore, the red signals inside the hotspot regions are weaker than the red signals outside. All evidence above strongly supports that the hot spots are indeed false-positive red signals, they should be the too strong green signals that pass the red filter. Their intensity is thus weakened by red optical filtering.

In order to identify and filter all hot spot signals, I merged all the samples using NLRS (Table 2.1), then observed the merged sample under IGV according to the position of the NLRS motif on the reference sequence. After checking the enriched signals at strange positions, we found that most of them have a distance with NLRS sites within around 300 bp. I then counted the number of signals in 300 bp windows, and the bins with hot spots enriched as candidate bins. After several investigations, I set a cutoff as 21, which includes most hotspot regions, and not remove the actual ORM signals.

Thus, I took 21 as a cutoff to picked out all 300 bp bins with more than 21 red signals, and I took these bins as suspect bins, which contain the hot spots. After mapping the calling regions to NLRS sites, we found that 84.89% of suspect bins containing NLRS sites. The left part also has an NLRS site located in the adjacent bin with a distance smaller than 300 bp. Such kind of slight position shift may be caused by the mapping deviation or indels. Fig. 2.15 shows an example of hotspots identified within final hot spot regions with NLRS sites.

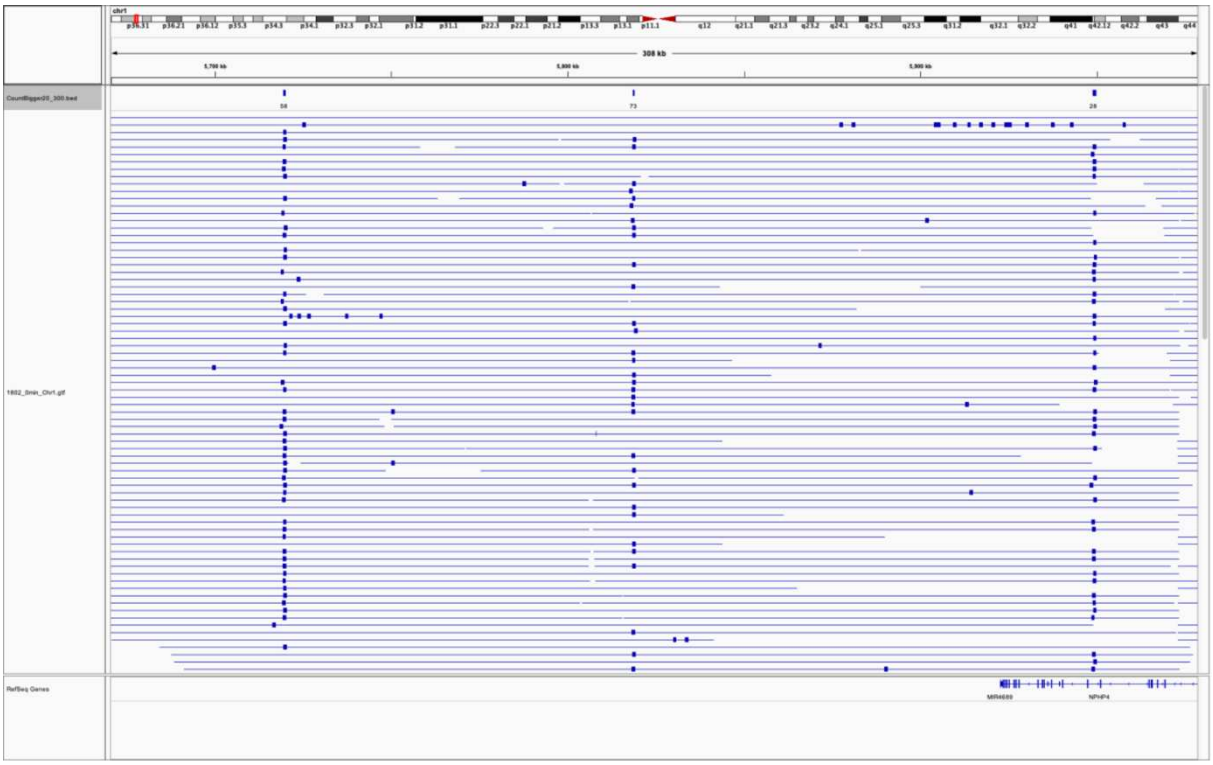


Figure 2.15. The obvious hot spot enrichment in calling suspect regions. The first line shows the three positions called as final hot spot regions. The number below them are 58, 73 and 26, respectively, which means how many signals within these hot spot regions. The fibers and ORM signals are visualized on IGV as in Fig. 2.9.

By this, we filter most false-positive red signals at hot spot regions, and the detailed statistics are shown in Table 2.2.

Table 2.2 Hot spot signals identified in different DLRS samples.

Dataset	Time after enrty of S phase	The total number of red signals	Number of filtered hotspost	Filtering percentage
C	0 min	1453750	67966	4.68%
	5 min	540882	23939	4.48%
	10 min	584523	9234	1.58%
D	0 min	877185	31544	3.60%
	20 min	1130340	33124	2.93%
	30 min	967749	22519	2.33%
	45 min	1371861	42572	3.10%
	60 min	1301848	46849	3.60%
	90 min	1582179	51780	3.27%

2.7 Segmentation for ORM labeling signals

Due to the low labeling efficiency of fluorescent dUTP used in our ORM approach (see section 2.8.6 for detail), ORM red signals detected from the Bionano platform are dispersed, and what we need is to identify replication origin regions. So, the idea is to cluster neighborhood ORM signals to form the continuous ORM segments. But how to set a proper cutoff value to distinguish the ORM signals revealing continuous regions or technical noise data. Firstly, we calculated the distance between all adjacent ORM signals and drew the distance distribution. Our working hypnosis is that the mixed distribution resulted from the real labeling signals and noise data, and we believe that, in general, both signal and noise follow a gaussian distribution. So, we introduced GMM (Gaussian mixture model) to classify the adjacent signal distance distribution into 2 gaussian distributions at first and took the intersection point of 2 distributions as the cutoff value (data not shown). But we soon discovered, the cutoff value varies a lot in different experiments, and sometimes the 2 classified distributions were very close so that even taking the cutoff value still leads to high false positives and true negatives that are hard to ignore. All of these may be due to the fact that part of the signal is at the junction of the real signals and the noise signals. In response to this situation, we decomposed the original distribution into 3 Gaussian distributions, and after improvement, we found that in almost all experiments, the tail position of the first Gaussian distribution is quite robust in our experimental setting (Fig. 2.16).

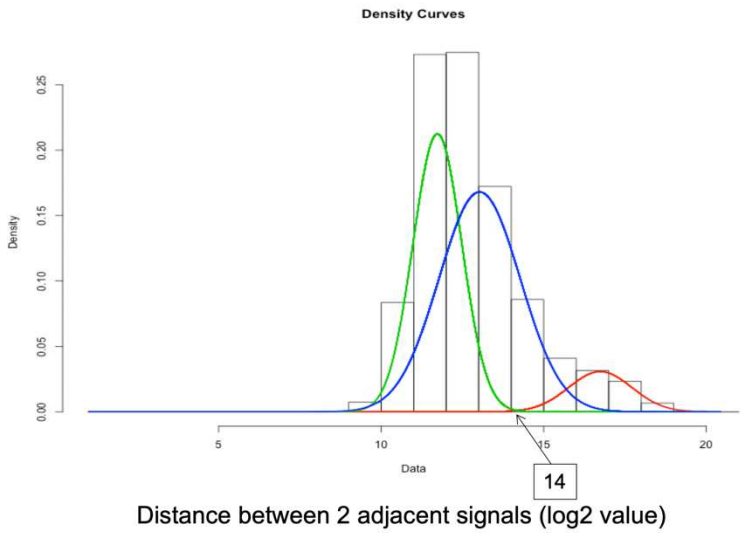


Figure 2.16. Three-way GMM for the distance between adjacent ORM signals. The x-axis gives the distance of adjacent signals after log2 transformation. The green, red and blue distribution represents the real signal, the noise signal, and the signal at the border of the two, respectively.

The final default cutoff was, therefore, set as 2^{14} (equals 16,384) bp. In this way, the primary ORM signal clustered tracks were generated. Considering the low labeling efficiency, the primary clustering may still not be able to get the complete segmentation representing a replication fork or an origin. So, for clustering the sub-segments, further clustering based on the primary ORM tracks is then necessary. Again, I calculated the distance between adjacent primary ORM tracks and introduced GMM to decompose the distance distribution of adjacent tracks into 3 Gaussian distributions (Fig. 2.17).

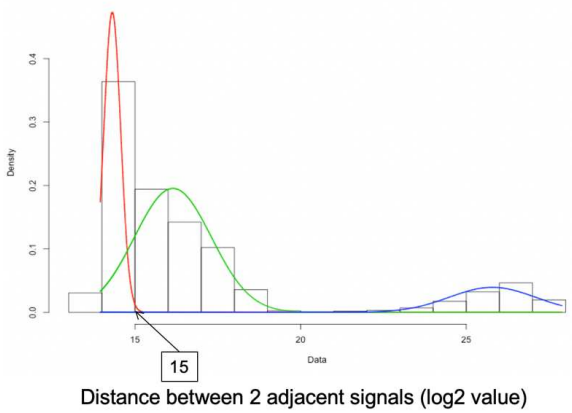


Figure 2.17. Three-way GMM for adjacent ORM primary tracks. The x-axis shows the distance of adjacent tracks after log2 transformation. The red, green, and blue distribution represent the distance between replication forks, the origins, and the noise, respectively.

It is worth noting that the adjacent distance distribution of primary ORM tracks shown in Fig. 2.17 is clearly distributed into two peaks. Why did we not just use directly the tail of the first peaks? This is because that the first peak obviously does not conform to a normal distribution, and in some experimental results, the second peaks cannot be observed (data not shown). Even if it can be observed, the tail of the first peak will be various in different experiments. We need to be cautious that the purpose of secondary clustering is to merge only the adjacent tracks representing the replication forks or sub-segments into a complete origin area, rather than merging most of the adjacent tracks representing origins into the broad initial zone. So, we still decompose the adjacent distance distribution into 3 Gaussian distributions and get an almost uniform first tail of the 1st Gaussian distribution as shown in Fig. 2.17. Based on this, we get the cutoff for the primary ORM track clustering: 2^{15} (equals 32,768) bp. The final segments that we got after two rounds of clustering have a good concordance with the published replication timing data, as well as the replication origins revealed by OK-seq data (Petryk et al., 2016) (Fig. 2.18). And the segments were also clustered around the genes related to DNA replication like Top1. All of these provide supportive evidence to the reliability of our ORM method.

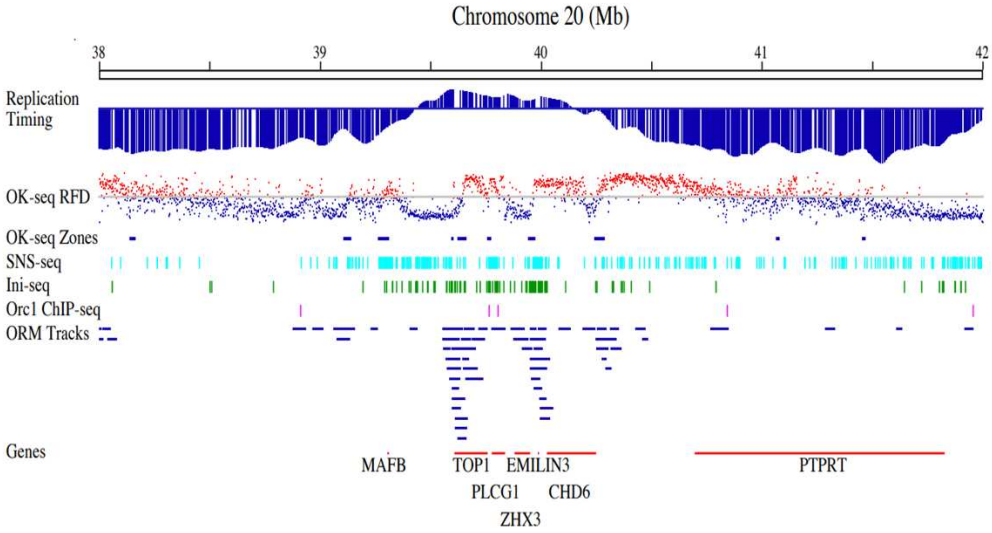


Figure 2.18. Good concordance between various data (Wang et al., 2020)**.** The first line is the replication timing curve. The right-side arrow shows the correspondence between the temporal order and the Replication Timing value.

The second line is the RFD curve obtained by OK-seq as in Fig. 1.13. The regions marked by dashed lines are replication origins called OK-seq. The third line is the ORM segment, and the last line is the gene positions, including Top1 with known replication origins around its promoter regions.

2.8 The reliability test for ORM segmentation

2.8.1 Track the trajectory of separated replication forks

For further testing that most ORM segments we got from the 0' samples are real replication origins, a series of experiments were designed and implemented. We hypothesized that, if the detected ORM tracks corresponding to replication origins at 0' data, with the increasing of waiting timing between the cells entering the S phase and the Fluo-dUTP incorporation, the distance between the two on-going replication forks from the same replication origins should become larger and larger even turn into 2 segments with a gap between 2 labeled replication forks.

We thus performed the experiments by incorporating the Fluo-dUTP by waiting 5 min, 10 min, 20 min, 30 min, 45 min, 60 min, 90 min after synchronized cells entering into the S phase. By introducing a longer waiting time, we expected to see a time-dependent movement of incorporation tracks on both sides away from the initiation zones. When we aggregated ORM replication tracks around early initiation zones (T-peaks), we indeed observed a time-dependent movement of incorporation tracks on both sides away from the initiation zones moving from single peak to double ones as expected (Fig. 2.19). In addition, the speed was estimated at a rate of approximately 1.65±0.31 kb per minute in agreement with the standard replication fork speed detected in human cells (1-3 kb/min) (Jackson and Pombo, 1998; Conti et al., 2007; Chagin et al., 2016). This evidence further supports the reliability of results obtained by the ORM technique.

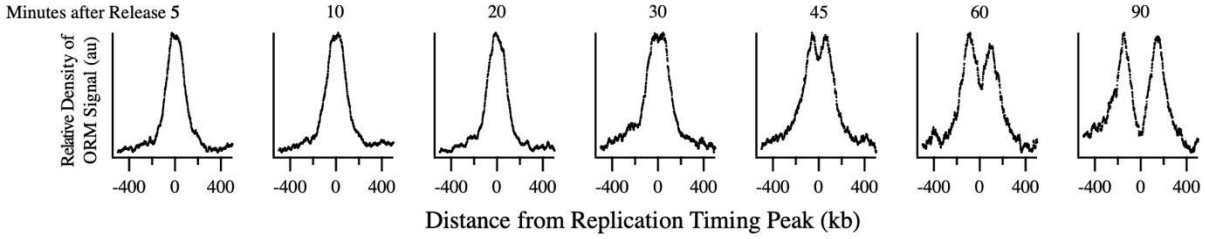


Figure 2.19. Time-dependent ORM signal movement around early initiation zones. 1,436 early T-peaks (data from UCSC Encode project), which replicate earlier than adjacent genomic regions, were used. We divided the nearby genome regions into a 1 kb window from downstream 500 kb to upstream 500 kb. We count the sum of signal number per window and get normalized fire efficiency by dividing the mapped fiber number in the corresponding bin to observe the ORM track enrichment changes around these regions in a different time, i.e. 0 min, 20 min, 45 min, 60 min, and 90 min after entry of S phase. In this way, we can calculate all fire efficiency values in +/-500 kb bins around 1,436 early T-peaks. The Y value will be the average normalized fire efficiency of all T-peak regions. The X value records the distance to the T-peak centers. The 0 represents the T-peak center position and +/- 400 represents 400 kb downstream and upstream region around the T-peak. The number on top of each distribution gives the minute after entry of the S phase.

2.8.2 The unexpected length distribution in all datasets.

Further digging on these additional experiments by labeling at different time points after the cells enter the S phase if these two segments are replicated forks separated from the same origin, the average length of segments before 30 minutes should be double than the segmentation length in 90 minutes, since the former is associated with two yet un-separated replication forks (considering the cutoff used in our analysis and the average replication fork speed is 1-3 kb/min) of the same origins, and the latter is more likely associated with each individual replication fork (the distance of on-going replication fork is large enough to be detected separately). We then checked the length of ORM tracks in samples with different waiting-time before labeling. We expect to see the average length of ORM segments representing replication origins in the earlier sample (e.g. 0 min sample) could be larger (even double) than the segment length in later samples, e.g. 45 min, 60 min, and 90 min samples corresponding instead to one of the two replication forks. Surprisingly, the length distributions of the ORM segments in 0 min is only slightly longer than the other samples.

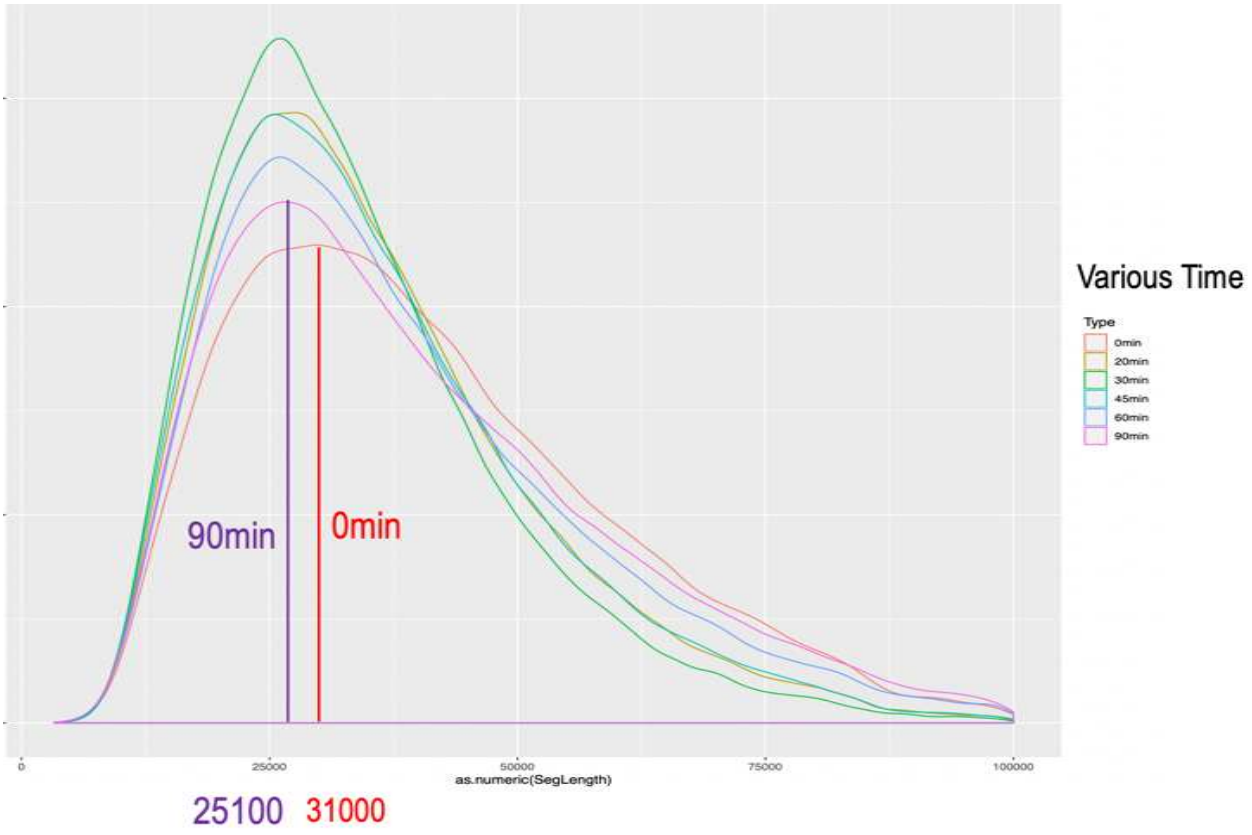


Figure 2.20. The ORM segment length distribution in groups of different time (0 min – 90 min) since the entry of the S phase. The red one is a 0 min distribution whose peak is around 30 kb, and most of the other peaks are around 25 kb.

2.8.3 Two hypnosis for explaining the unexpected length distribution

To explain such observation, three models will be possible. Fig. 2.21 propose several kinds of the possible model caused by biological reason and technical bias. In the first model, numerous newly fired replication origins might mix with single replication forks in 90 min dataset, which increases the average ORM segment length of the overall distribution. Or in a second model, one of the two replication forks from the same replication origin failed to restart in a 0 min dataset after the release of synchronization. And in the last model, it's due to the low labeling efficiency in incorporating florescent dUTP in the current ORM approach.

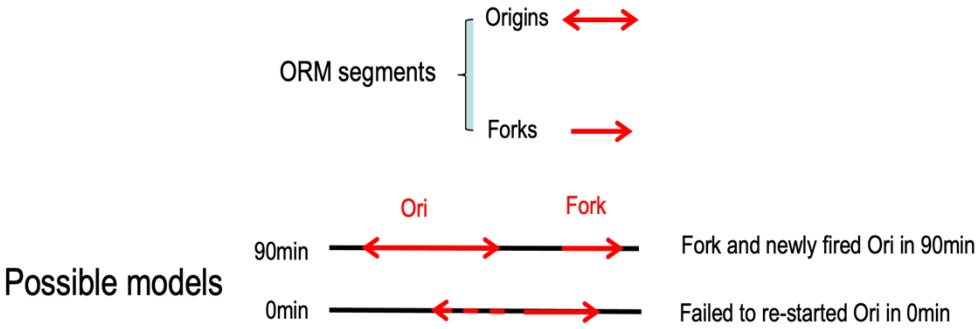


Figure 2.21. Schematic of models. The red solid part represents incorporation segments; the dashed part represents failed restarted replication forks or unlabeled segments, and the arrow shows the direction of replication forks.

To further distinguish these models, in the first model hypnosis, we guess the number of fibers with 2 closed adjacent segments should increase after 30 min when the signal distribution showing the peak split trend. So, I checked the number of segments per fiber to see whether, in 90 min sample, we can detect more fibers containing 2 ORM segments. And again, the percentages of fibers containing 2 or more segments per fiber are similar to each other in all datasets (Fig. 2.22), indicating that the unexpected length distribution may be caused by the failure in the detection of one of two replication forks.

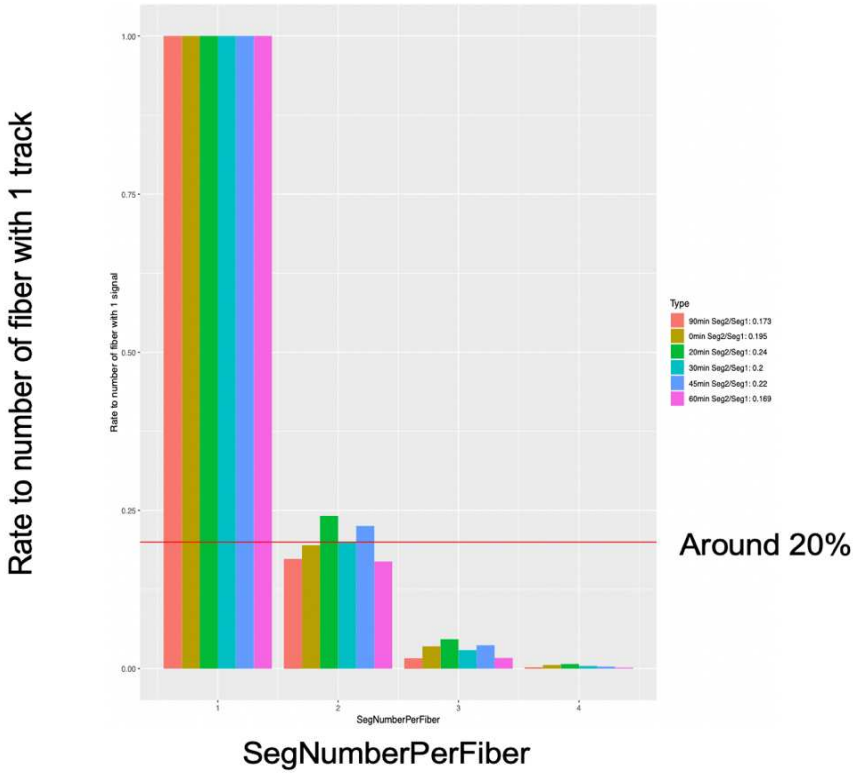


Figure 2.22. The relative number of fibers containing 1-4 ORM segments. The amount of fiber-containing 1 ORM segment was normalized to 1 and the relative number of fibers containing 2, 3, or 4 ORM segments were shown. All datasets show a similar distribution.

All of this evidence suggested that the ORM segments obtained in the current analysis might detect only one of the two replication forks from the same replication origin, and the enrichment around T-peaks, in fact, are piled up by single replication forks as shown in Fig 2.23. The analysis of replication fork directionality of the ORM segments further confirmed this hypothesis (see Chapter 4 for detail).

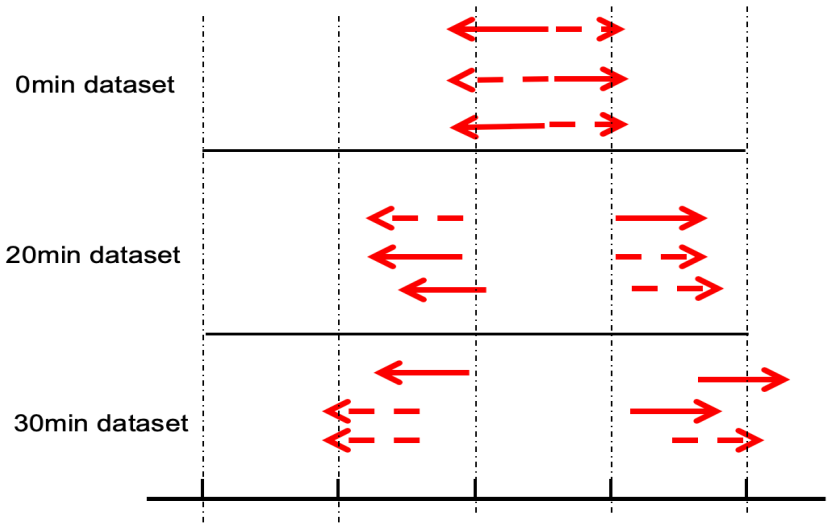


Figure 2.23. Schematic diagram to explain the ORM segment enrichment around T-peaks. The solid flashes show the replication forks after the synchronization release detected by ORM segments, and the broken flashes indicate the replication forks failed to be detected by ORM segments.

2.8.4 Verification of potential model

In order to further confirm the hypothesis model, our collaborators Karel Proesmans and John Bechhoefer (Simon Fraser University, Canada) performed rigorous probability calculations to quantify the possibility of the incomplete Labeling, which I detailly described below.

Under rather general assumptions, one expects the number of segments on a fiber of length *l* to be Poisson distributed:

$$p(n|l) = e^{-\lambda}\frac{\lambda^n}{n!}, \tag{2.3}$$

Where λ is proportional to the length:

$$\lambda = c_0 l. \tag{2.4}$$

To test this hypothesis (and to find c0), one can check that

$$\frac{p(n+1|l)(n+1)}{p(n|l)l} = c_0, \tag{2.5}$$

By substituting our data operation in 0-90 minutes, we get

$$c_0 = (1.3 \pm 0.1) \cdot 10^{-4}. \tag{2.6}$$

Due to Bayes theorem, we have

$$p(l|n) = \frac{p(n|l)p(l)}{p(n)} = e^{-c_0 l}\frac{(c_0 l)^n}{n!}\frac{p(l)}{p(n)}. \tag{2.7}$$

Therefore, p(l|n) / l^n is, up to a constant, independent of n,

$$p(l|n)/l^n = c(n)p(l). \tag{2.8}$$

Which could be implies

$$p(l|n)l^{-n} \sim e^{-c_1 l} \tag{2.9}$$

Again, substitute segments which signal number equal to 4, 6, 8, 10, 12 and corresponding length, we got Figure 2.24.

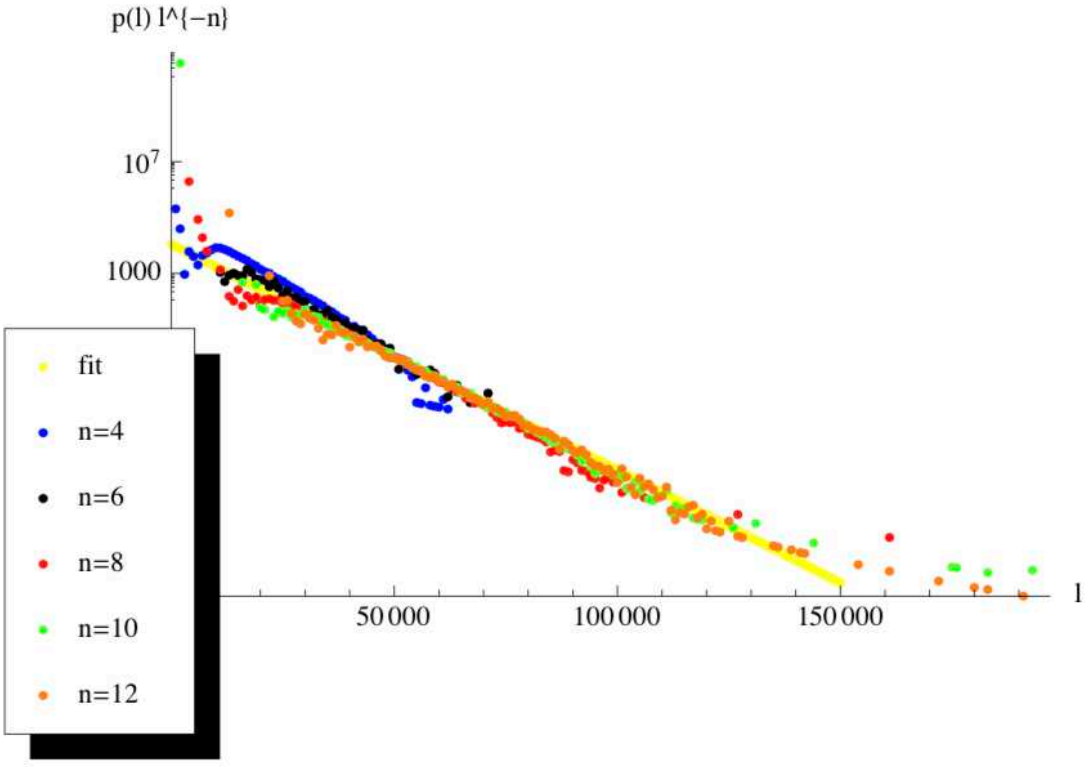


Figure 2.24. Comparison of p(lln) / ln in function of (1) for n = 4 (blue) 6 (black), 8 (red), 10 (green) and 12 (orange). In above plot, we found no matter which kind of color has a good consistence with fit line, which means the correspondence of this kind of function between segments number and fiber length exists.

Where a fit give

$$c_1 = (1.9 \pm 0.1) \cdot 10^{-4}. \tag{2.10}$$

Then, proper normalization gives:

$$p(l|n) = \frac{c_1^{1+n}}{n!} l^n e^{-c_1 l}. \tag{2.11}$$

We are now ready to determine the separate probabilities *p(n)* and *p(l)* (i.e., the probability that there are n signals on a segment and the probability that the length of a segments is equal to l). From Bayes theorem, we know that

$$\frac{p(n)}{p(l)} = \frac{p(n|l)}{p(l|n)} = \frac{e^{-(c_1-c_0)l} c_0^n}{c_1^{n+1}}, \tag{2.12}$$

And as p(n) should be independent of l and vice versa, this leads to

$$p(n) = \frac{c_1 - c_0}{c_1} \left(\frac{c_0}{c_1}\right)^n \tag{2.13}$$

$$p(l) = \frac{1}{c_1 - c_0} e^{-(c_1-c_0)l}. \tag{2.14}$$

So, we can draw the probability curve shown in Fig. 2.25.

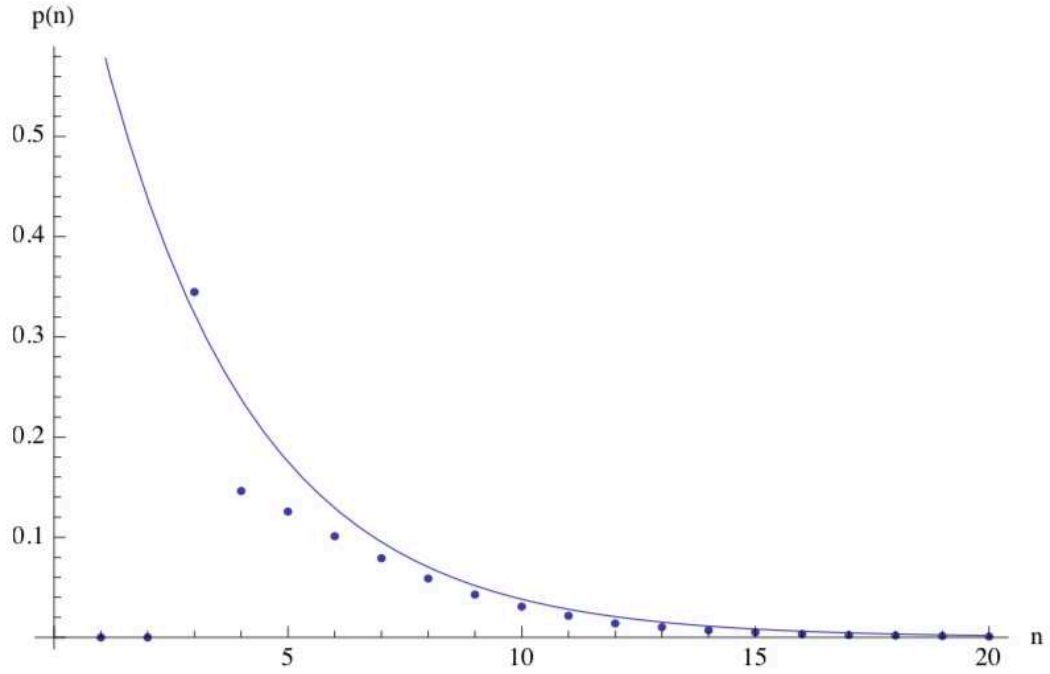


Figure 2.25. *p(n)* in the function of *n*. It is easy to see the high probability of segment with only 1 or 2 signals cannot be ignored. The incomplete labeling may hide the missing part of replication forks in the third model that we proposed.

2.8.5 Regaining the neglected signals

2.8.5.1 Too many filtering signals by GMM

Previously, because of the introduction of GMM (Gaussian mixed distribution) segmentation, a huge number of single signals are filtered by the segmentation clustering (we only kept the segments with at least 3 signals), and the analysis in the previous section showed that there are large numbers of incomplete labeling tracks containing few ORM signals (i.e. 1 or 2 signals). If so, how can we include all detected ORM signals in our analysis? The answer is using all signals enrichment to call initial zone instead of ORM segments organized by clustered signal. In the previous GMM segmentation process, only ORM tracks containing at least 3 clustered signals would be recognized as an ORM segment. And such kinds of fibers only account a small fraction (33.63%) of all fibers (Fig. 2.26). It means that we can increase around 3-fold statistic to call initial zones, if we add the ones containing fewer ORM signals.

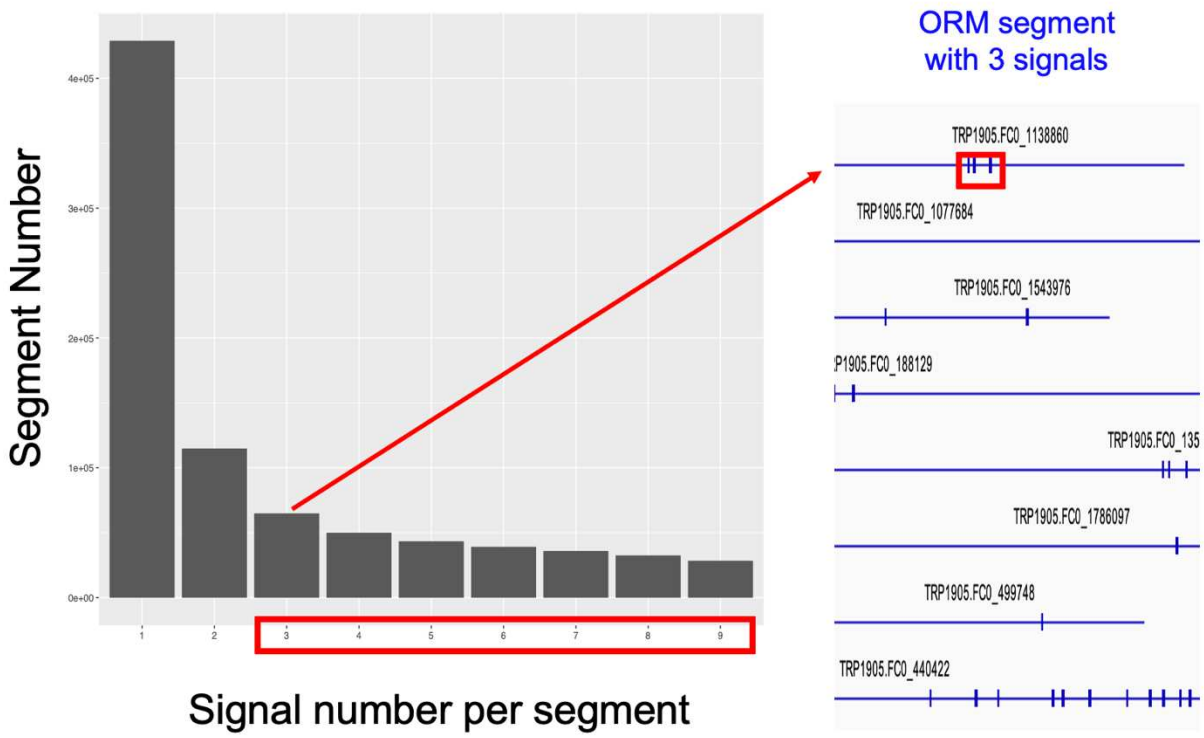


Figure 2.26. The histogram of segment number containing 1-9 ORM signals. There are large numbers of fiber-containing only 1 ORM red signal, and also a lot of segments with only 2 signals. The red frame includes all segments in the previous analysis, and the total number of the segment with at least 3 signals even much smaller than the number of segments with only 1 signal. The right part is a screenshot for DNA fibers with signals under IGV as in Fig. 2.9, and the red arrow shows an example of the segment containing 3 clustered ORM red signals.

2.8.5.2 The filtering signals can really represent incomplete labeling segments

Before we include these isolated labeling signals, we still need to check whether the sparse signals detected by our ORM technique are indeed real signals related to DNA replication initiation instead of technical noise. I picked out all fibers with a single signal and observed such kind of single signals also follow the time-dependent enrichment movement around T-peak regions in 0 min, 20 min, 60 min (Fig. 2.27), as the other ORM signals within ORM tracks (with at least 3 signals) shown in the previous section (Fig. 2.19).

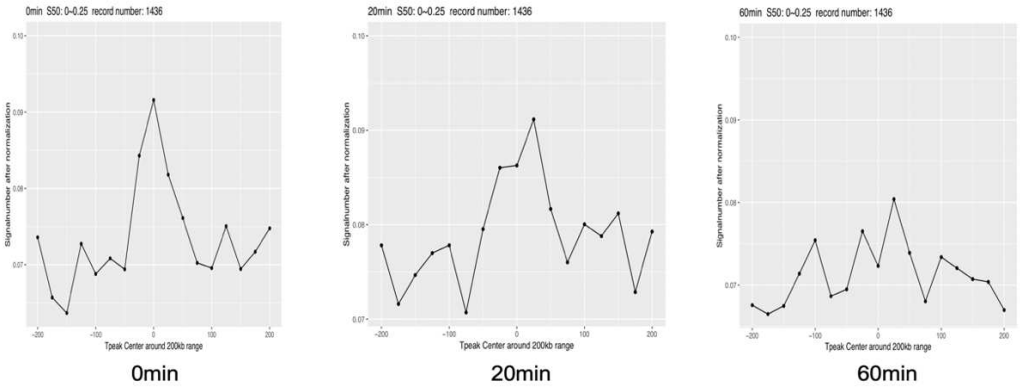


Figure 2.27. Single signal enrichment around T-peak centers. The single signal enrichment also consists of the time-dependent enrichment split by different waiting times after the release of the S phase.

Therefore, we can conclude that almost all ORM signals are real biological signals associated with DNA replication programs instead of technical noise, and the labeling efficiency in the current ORM approach is really quite low leading to very sparse labeling.

2.8.6 The explanation for sparse labeling

The reason why dUTP can't fully mark ongoing replication regions could be dUTP concentration is too low to perform an efficiency label. To check the influence of cell synchronization on the ORM segment length, I checked the ORM segments detected in the asynchronous sample and observed the replication tracks average 23.9±35.5 kb in length in the HeLa data and 27.5±40.4 kb in length in the H9 data, which is comparable to the length of tracks in the synchronized data (Fig 2.28). It highly suggests that it's not a problem that results from the cell synchronization. We, therefore, concluded that the sparse labeling is due to lower labeling efficiency.

Therefore, we increased the labeling dUTP concentration to introduce to asynchronous cells to see if we can increase labeling rate than previous concentration or see more fibers with 2 segmentations than 0 min synchronized data. Surprisingly we still didn't find any obvious changes (Fig. 2.29).

This unexpected result illustrated that there must be some technical or biological limitation in the current experimental setting, which limits the florescent dUTP incorporated within ongoing replication forks. Unfortunately, temporarily, we can't label all ongoing replication regions by increasing dUTP concentration.

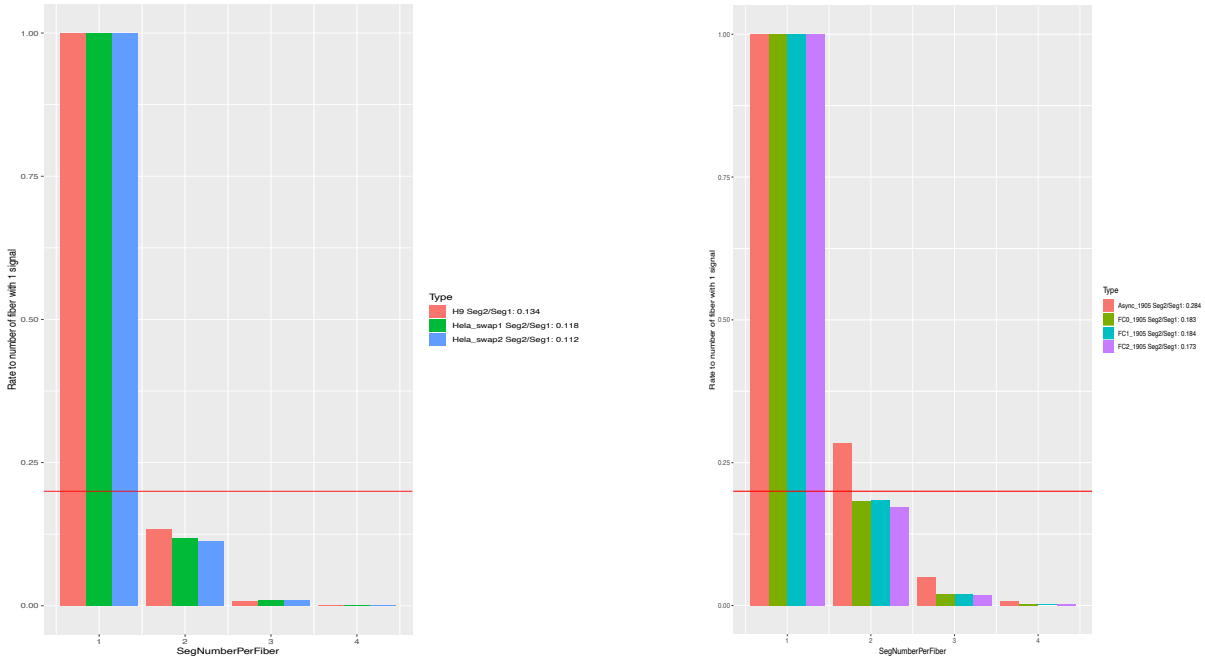


Figure 2.28. The relative number of fibers containing 1-4 ORM segments from ORM data of asynchronous (left) and synchronous (right) cells. The amount of fiber-containing 1 ORM segment was normalized to 1 and the relative number of fibers containing 2, 3, or 4 ORM segments were shown.

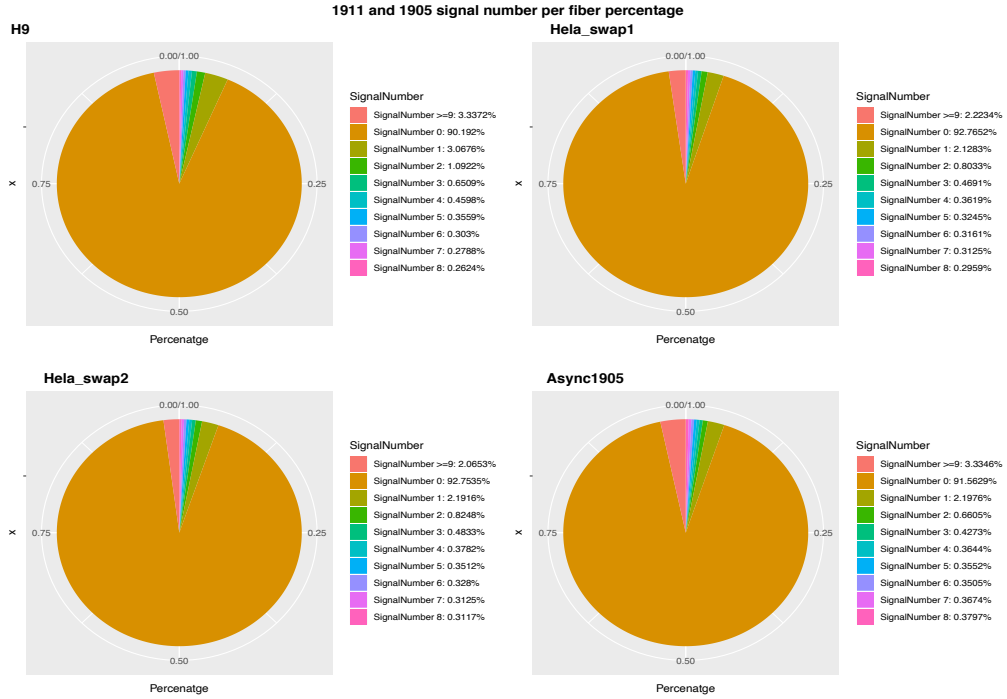


Figure 2.29. Signal number per fiber percentage in 3 folds concentration labeling experiments and normal concentration. H9 is an experiment with 3 folds concentrations in the H9 cell line, Hela_swap1 and Hela_swap2 are two replicates with 3 folds concentrations in the Hela cell line. And async1905 is the experiment with a normal concentration in the Hela cell line.

CHAPTER 3

Replication initial zone calling

3.1 Calculation of normalized ORM signal density

As described in Chapter 2, most detected ORM signals are indeed biological signals and not technical noise, and the ORM signals from the 0 min samples correspond to replication initiation events. I, therefore, used all ORM signals of 0 min samples to perform the initiation zone calling. The entire genome is divided into 1 kb bins, and the normalized signal density was calculated for each bin (Fig. 3.1).

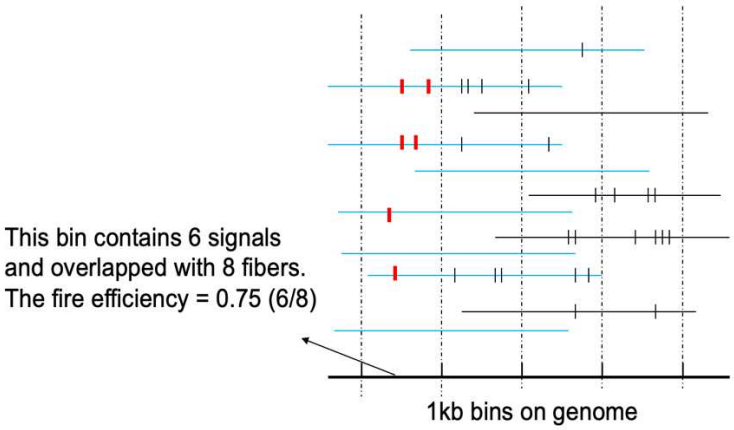


Figure 3.1. The schematic diagram for normalized signal density. The dashed line means the edge of bins. Within each bin, normalized signal density = signal number within bin / overlapped-fiber number. The arrow shows an example bin in the calculation of signal density (or fire efficiency in case of 0' data), all ORM signals in this bin were marked by red and overlapped fibers marked by blue color.

The signal number can reflect the fire probability within each bin. Meanwhile, the DNA fiber coverage distribution in the Hela cell line has an obvious aneuploid character. We classify all genomes into haploid, diploid, triploid and tetraploid regions of the genome. The maxima fiber coverage in these 4 regions are about 376x (haploid), 705x (diploid), 1,047x (triploid), and 1,400x (tetraploid). Such multiple differences are consistent with polyploidy characteristics. On average, the variation of fiber coverage distribution is not huge. Considering the variation of DNA fiber's coverage within each bin, which will cause the bias of signal number, the signal density needs to be normalized by dividing the overlapped DNA fibers number for each bin. After several trials, it was found that the number of signals contained in one 1 kb bin is very limited. For one ORM sample, more than 80% of the bins have less than 2 signals with many empty bins without any signals, and the average number of overlapped fibers in one 0 min dataset can reach more than 240. This leads to the normalized signal density generally approaching zero. In order to increase the number of signals containing in the bin without reducing the resolution, 10 kb bin was thus finally selected, which slides forward on the entire genome in steps of 1 kb. In order to be able to better visualize on the genome browser, such as IGV, the calculated signal density was recorded as the 1 kb length bins around the centers, for example, the 10 kb bin region from 1000 ~ 11000 has a

normalized signal density value 0.0037 will be recorded as 1 kb bin region 6000~7000 with this normalized signal density value.

3.2 Normalized signal density smoothing

When normalized signal density is calculated, we get the primary fire probability distribution. However, this raw data is very noisy to call peaks reflecting the real initial zones. So, smoothing is necessary before the peak calling. In the beginning, I tried to apply the Gauss fitting to our data, but it will change the shape of the raw distribution, such as plateaus, cliffs, and asymmetric hills into standard bell shape (data not shown). This will lead to a large amount of information loss, and the result of peak calling will be prolonged, shortened, or shifted to one side.

So, we finally adopt LOESS fitting to do the smoothing operation. It is more accurate to fit the successive area as long as possible, but if the total length of the chromosome is used as the unit, it will consume huge computing resources. Considering the calculation time and memory size, I cut the entire genome into fractions of 160 kb window to perform the LOESS fitting within each window. But I soon discovered that no matter how I adjust the parameters, a smooth simulation curve that reflects the original shape can be obtained in a single window, but there will be a very obvious gap at the junction of the adjacent windows (Fig. 3.3).

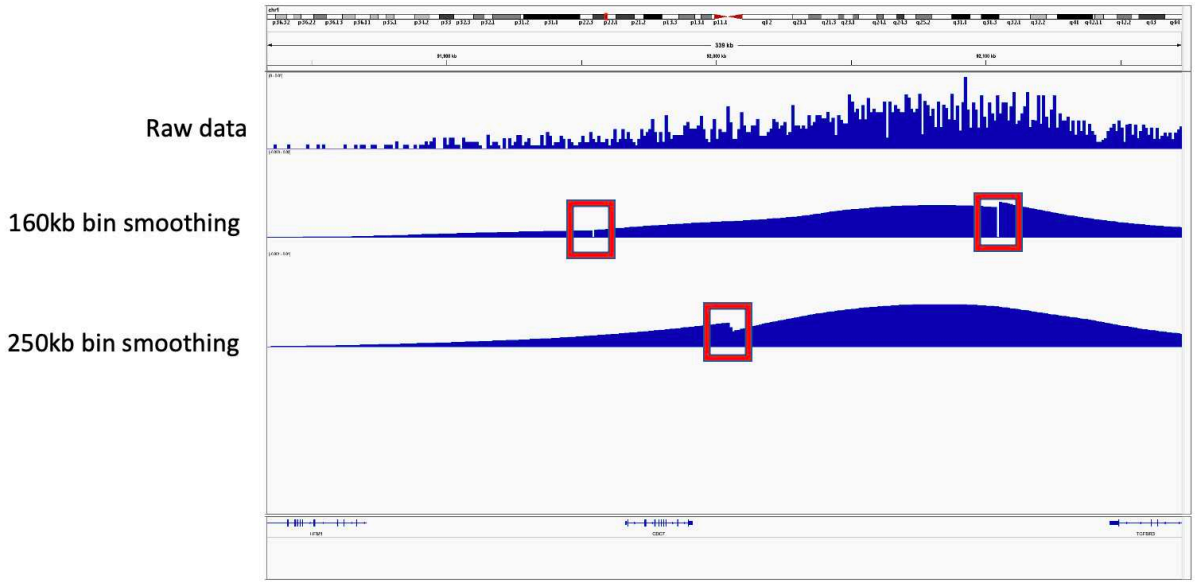


Figure 3.3. Gaps caused by LOESS fitting in adjacent windows. The raw ORM signal intensity in 10 kb sliding windows (with 1 kb step) is shown together with the LOESS fitting within 160 kb or 250 kb bins. The red rectangles show the gaps in junction between two adjacent windows.

In order to solve this problem, an 8 kb overlapping transition area was set up to avoid discontinuity of fitting at the junction of adjacent 160 kb windows. Within the 8 kb overlapping regions, the final smoothed values were the averages of values from LOESS fitting of adjacent windows, weighted by their distance to the corresponding window (Formula 3.1). For example, for a given 8 kb transition area, the weight for the 8 values in 8 kb of the transition area in the previous 160 kb window will be a sequence with a step size of 0.125, which is decreased from 1 to 0 accordingly.

And the weight for the 8 values in the later 160 kb window will be a sequence with a step size of 0.125, which is increased from 0 to 1 accordingly. Let's call the 8 smoothing density values in 8 kb transition area of previous 160 kb window PV1~PV8 and the weight for the PV1~PV8 as PW1~PW8. The PW1~PW8 will be 0, 0.125, 0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1. Similarly, the 8 values in 8kb transition of the later 160kb window will be recorded as LV1~LV8 and their weight will be called LW1~LW8 (1, 0.875, 0.75, 0.625, 0.5, 0.375, 0.25, 0.125, 0). The final values of the 8 kb bin will be FV1~FV8. The *n*th final value FVn (where n=1 to 8) in the transition area is calculated by the formula below.

$$FVn = PWn \times PVn + LWn \times LVn \qquad (3.1)$$

In this way, a smoothing fitting signal density distribution reflecting the original shape information was obtained (Fig. 3.4).

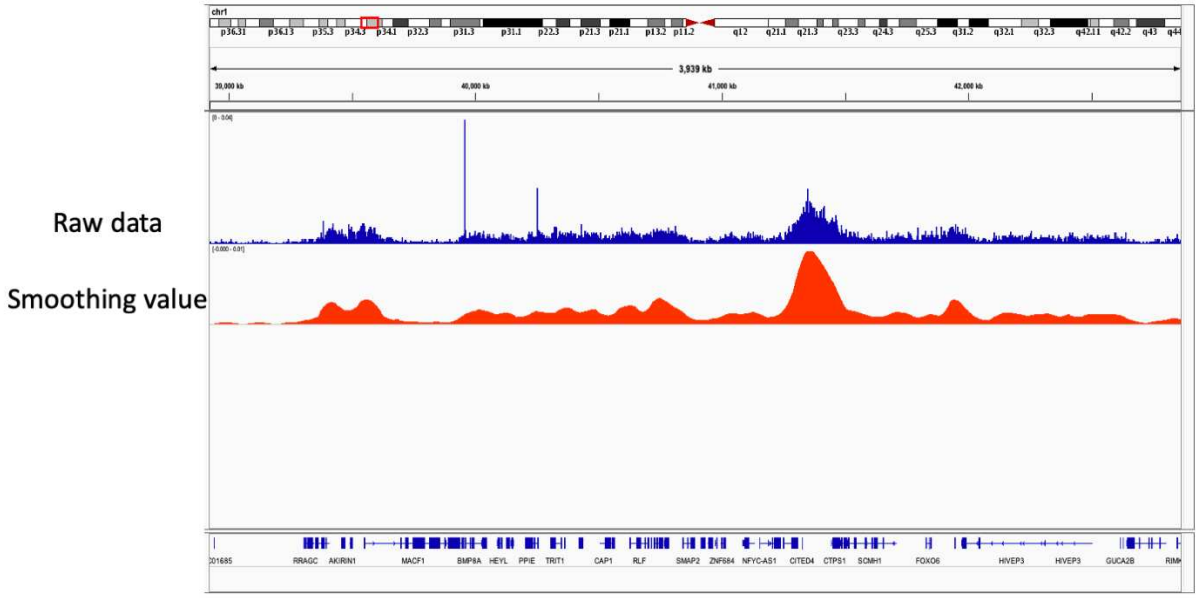


Figure 3.4. Smoothing LOESS fitting of ORM data. The blue distribution is the raw normalized signal density in 10 kb sliding windows (with 1 kb step), and the red one is the density profile of normalized signal density after LOES smoothing fitting (160 kb window) with the following parameters: (α = 0.75 and the polynomial degree = 2)

3.3 Peak area recognition

Because based on our ORM data and those obtained by other methods, such as OK-seq, there is no obvious fixed replication origin site in human cells (Fig. 2.18). We observed broad initiation zones consisting of many initiation sites instead (Fig. 2.18). Our final purpose is thus to detect the initial zone with high fire efficiency by doing ORM signals peak calling. Confronted with different shapes of normalized signal density distribution, the traditional peak calling result by tools like Macs can't meet the automatic recognition for initial zone length. And for most calling peaks by Macs2, even if using the option for broad peak calling, the results are just a few bins around the bin containing peak point (about 5 kb, data do not show). However, the resolution of the current ORM methods is estimated at ~15 kb (based on the labeling efficiency obtained in our probability

calculation, see section 2.8.4), and the ORM segment length is on average 19.5 kb. Concerned with the limitation of labeling efficiency, 5 kb initial zone is hard to cover all possible origin sites reflected by ORM signals.

Point to the problem, we have developed a new automatic peak recognition algorithm. We classified all bins into four categories: up, down, valley, and peak according to the changing trend of its normalized signal density value comparing with the corresponding value of neighboring bins. We got 26,196 bins with peak label (local maximum on the smooth ORM density profile) and we take the peak point to the two side closet bins with the "valley" sign as 26,196 primary peak areas. Then we need to further refine the core region inside the primary peak area.

3.4 Core region refining

3.4.1 The aggregated density percentage

In our experiment, the fluorescent dUTP has been transfected into the cells once and runs out very quickly. For the replication initiation zones, the ORM signals should be aggregated around the core regions. Therefore, on the density profile, the normalized signal density at the edge of the core area will drop rapidly. Thus, I calculated the sum of normalized signal density values in all primary peak areas and the density percentage of each bin to the corresponding density sum of the primary peak area (Fig. 3.5). Then, I averaged the percentages 50 kb upstream and downstream of all peaks to examine the ORM signal density around peak centers.

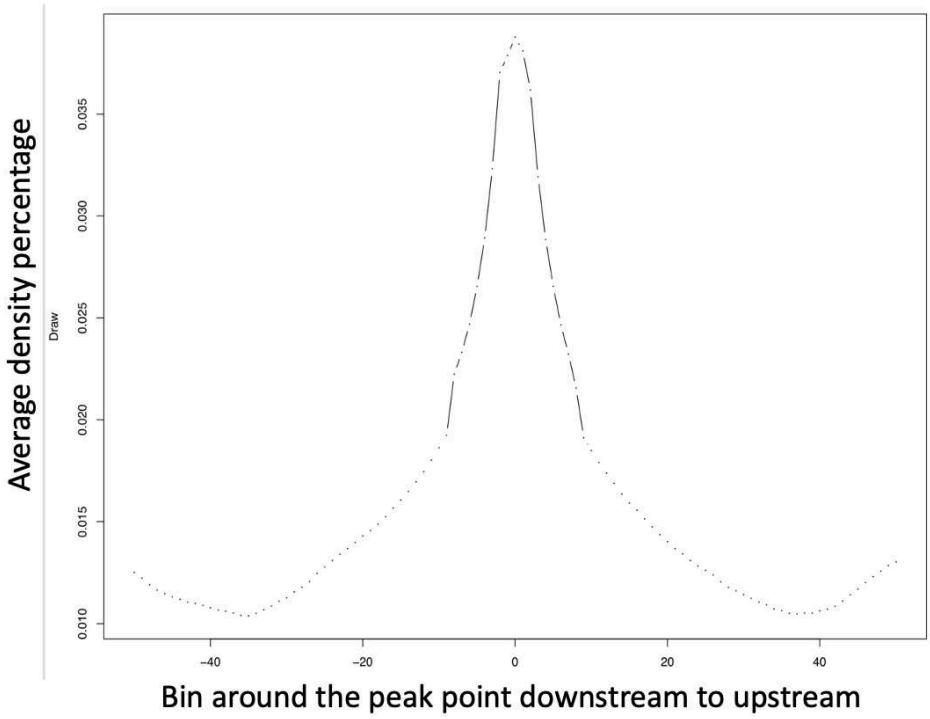


Figure 3.5. Average ORM density profile around centers of primary peak areas. The ORM density percentage decreases very quickly within 8 kb distance around peak centers.

The shape of the ORM signal density profile observed in Fig. 3.5 perfectly meets our hypothesis, and it indicates that further refining for the final initial zone is also necessary.

3.4.2 Estimate proper signal percentage cutoff to call core regions of initiation zones

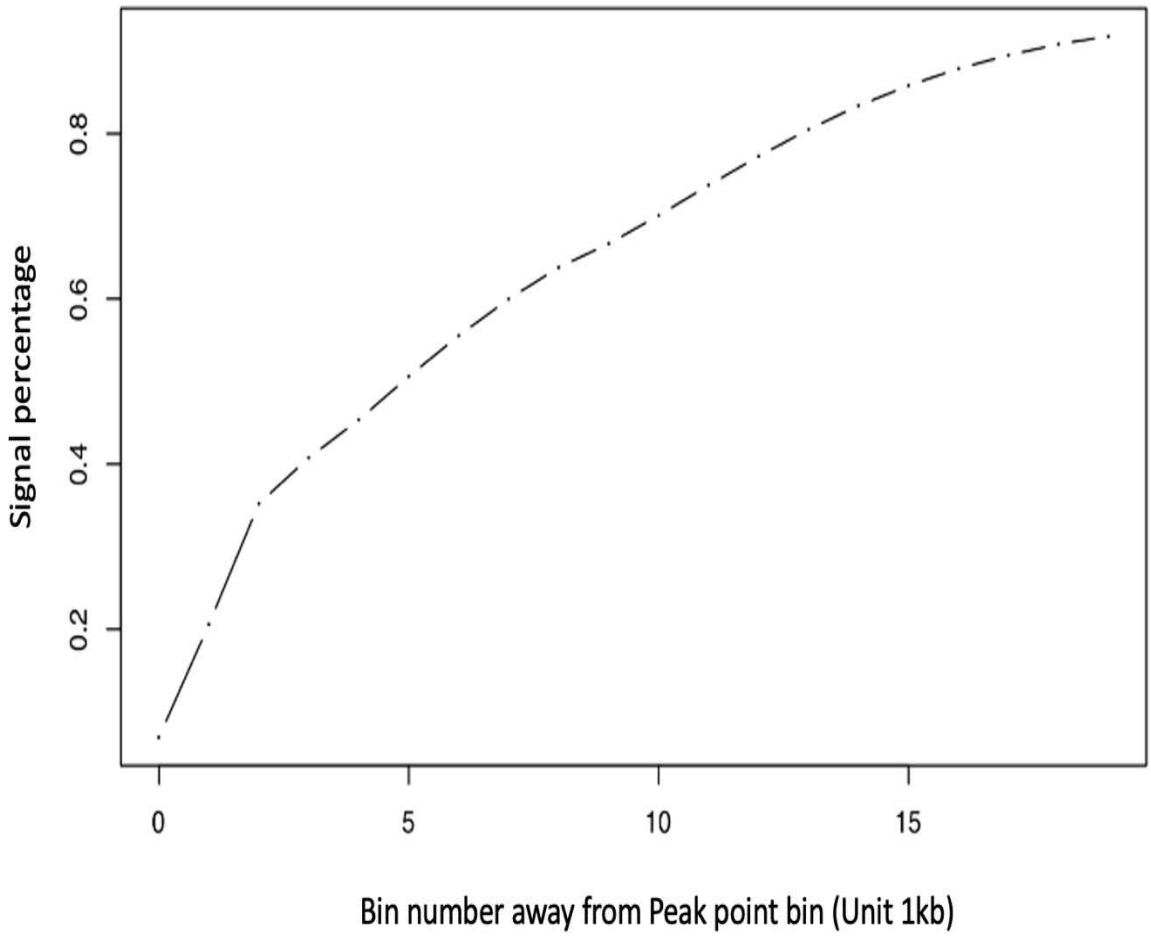


Figure 3.6. Cumulative signal percentage sum of merged 0 min dataset. The X value is the distance away from peak positions on both sides. Y value represents the average signal percentage sum of all the bin within the region from peak to the corresponding distance of X value on two sides.

In Fig 3.5, we observed an aggregated density percentage with a shape of a sharp peak. Let's call it ADPP (aggregated density percentage peak). Within +/-8 kb around the ADPP's center (corresponding x value = 0), the average density percentage value (Y value) decreased very quickly. It means most ORM signal clustering ranges from -8 kb to 8 kb. However, the +/- 8 kb range is just a result of average data. We can't apply the 8 kb around peak positions to all primary peak areas with different steepness. For the shape of the primary peak areas are more precipitous or flat, the +/- 8 kb range maybe extend or shrink. So, we prepared to use the percentage of signals as a standard to divide the core region. Firstly, the core region starts from the bin with peak positions (the bin with the label "peak") of primary peak areas. Every time extend 1 bin to each side and recalculate the accumulated signal percentage within the extended core region to the entire primary

peak area. Repeat this step until one side up to the edge of the primary area. Fig 3.6 reflects the relationship between accumulated percentage and the bin number away from the bin with peak position. In Figure 3.6, an obvious turning point shows up when signal percentage up to around 0.38~0.4.

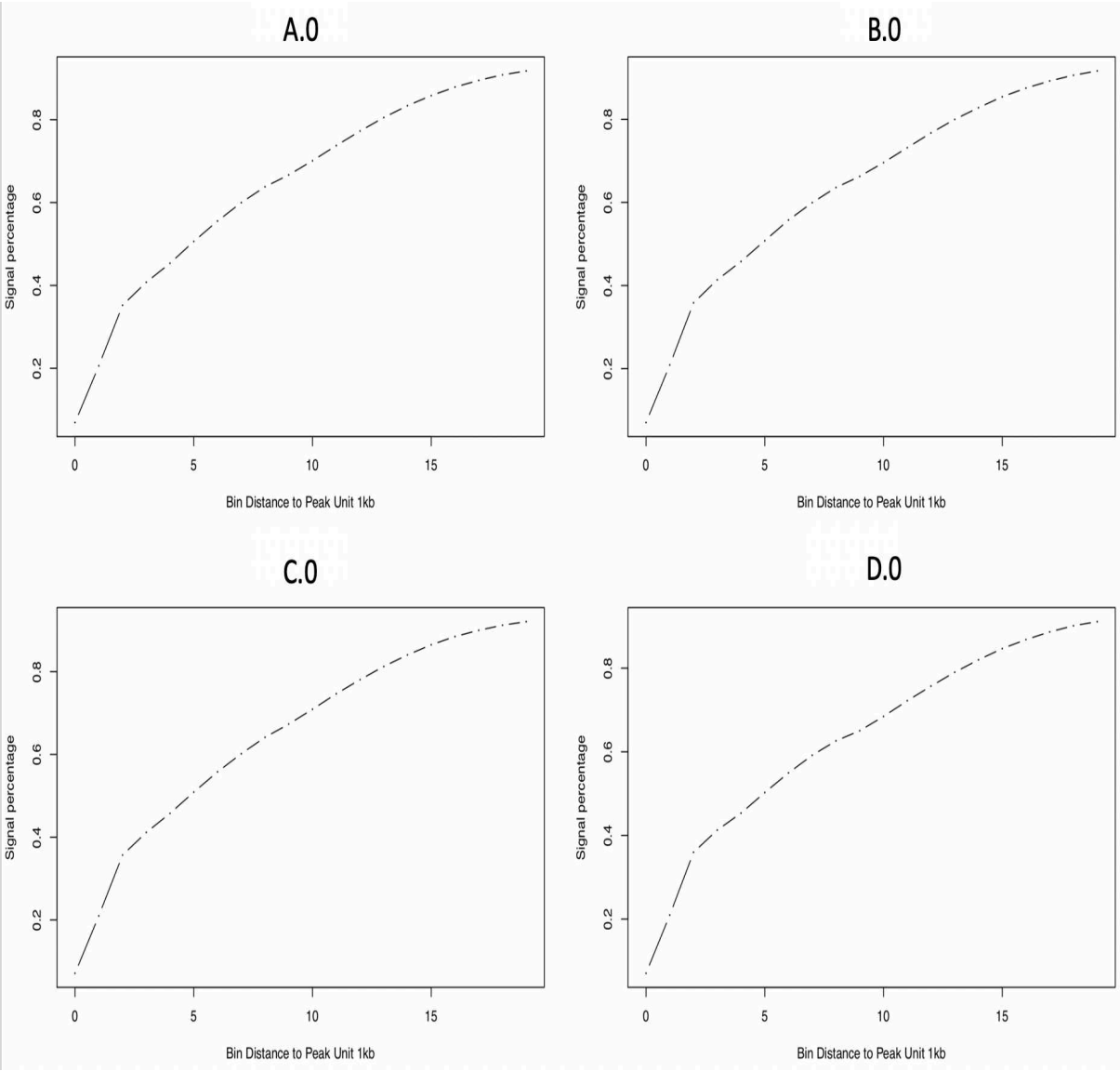


Fig 3.7. Cumulative signal percentage sum in each of the 4 replicates of 0' min ORM data. Figures show cumulative signal percentage sum around peak position as in Fig 3.6.

This process was performed on all 4 biological replicates of HeLa 0' data (A.0, B.0, C.0, and D.0 of table 2.1) as well as the combined dataset and got very similar results (Fig. 3.7). Based on figure 3.7, we set a conservatively 40% as a cutoff to divide the core regions from primary peak areas, because 40% is a little bit higher than the turning point in Y-axis. The core region starts from the bin where the peak is located and takes a single bin as the smallest unit to extend. Every time it compares the signal percentage values of the adjacent bin on the left side and right side of the core

region, and then include the bin on the side with a larger percentage value. Repeat this step until the accumulated signal percentage in the core region reaches 40% of the total signal in the entire primary peak area.

Then 5 files (4 replicates and merged dataset) record core regions with 1 kb resolution will be generated. As shown in Fig. 3.8, within this file, all information related to each bin has been recorded, including chromosome, bin start, bin end, normalized signal density, density percentage of corresponding primary area, the trend sign (Up, Down, Valley, peak in chapter 3.3) and adding column marked by 0/1 to represent whether one bin is contained by one core region (1 or 0 represents, respectively, whether it is belonging to core region bin or not).

chr1	133001	134000	0.038270.	013336	UP 0
chr1	134001	135000	0.038557	0.013436	Up 0
chr1	135001	136000	0.038837	0.013534	Up 0
chr1	136001	137000	0.039108	0.013628	Up 0
chr1	137001	138000	0.039368	0.013718	Up 0
chr1	138001	139000	0.039614	0.013804	Up 0
chr1	139001	140000	0.039847	0.013885	Up 0
chr1	140001	141000	0.040063	0.013961	Up 0
chr1	141001	142000	0.040261	0.01403	Up 1
chr1	142001	143000	0.040439	0.014092	Up 1
chr1	143001	144000	0.040596	0.014147	Up 1
chr1	144001	145000	0.04073	0.014193	Up 1
chr1	145001	146000	0.04084	0.014231	Up 1
chr1	146001	147000	0.040923	0.01426	Up 1
chr1	147001	148000	0.040977	0.014279	Up 1
chr1	148001	149000	0.041002	0.014288	Peak 1
chr1	149001	150000	0.040996	0.014286	Down 1
chr1	150001	151000	0.040956	0.014272	Down 1
chr1	151001	152000	0.040881	0.014246	Down 1
chr1	152001	153000	0.040773	0.014208	Down 1
chr1	153001	154000	0.040634	0.01416	Down 1
chr1	154001	155000	0.040465	0.014101	Down 1
chr1	155001	156000	0.040266	0.014032	Down 1
chr1	156001	157000	0.040039	0.013952	Down 0
chr1	157001	158000	0.039784	0.013863	Down 0
chr1	158001	159000	0.039501	0.013765	Down 0
chr1	159001	160000	0.039191	0.013657	Down 0

Figure 3.8. Example lines show the data format of temporary files related to identifying core regions of initiation zones detected on ORM signals density profile.

3.5 Filtering and initial zone calling

Through the operation above, we extracted the core regions from the primary peak areas. However, there may still be special waveforms caused by filtering, noise data, and superimposition of the fitting signals in the transition areas, which may cause false-positive peaks. In this regard, we still need to filter core regions to get the final initial zones.

3.5.1 Overlapped replicates number filtering

There are 4 biological replicates of Hela 0 min datasets (A, B, C, D) (Table 2.1), which were further combined into one merged 0 min dataset. In the merged dataset, normalized signal density = signals number sum of 4 replicates/overlapped fiber number sum of 4 replicates. The 40% calling protocol described above was implemented for each of them. Although the shapes of normalized signal peak distribution are generally similar (Pearson score is between 0.69~0.97) like the 5 ORM smooth density profiles shown in Fig 3.9, it displays also some variations amongst 4 replicates and the merged dataset, which results in the different core regions (Fig. 3.9). It's reasonable to suppose that the core regions robustly detected by most replicates show higher firing efficiency and have higher confidence than those only identified in one or two replicates. Therefore, we decided to only keep the core regions in the merged dataset, which overlapped with the core regions identified within at least 3 biological replicates.

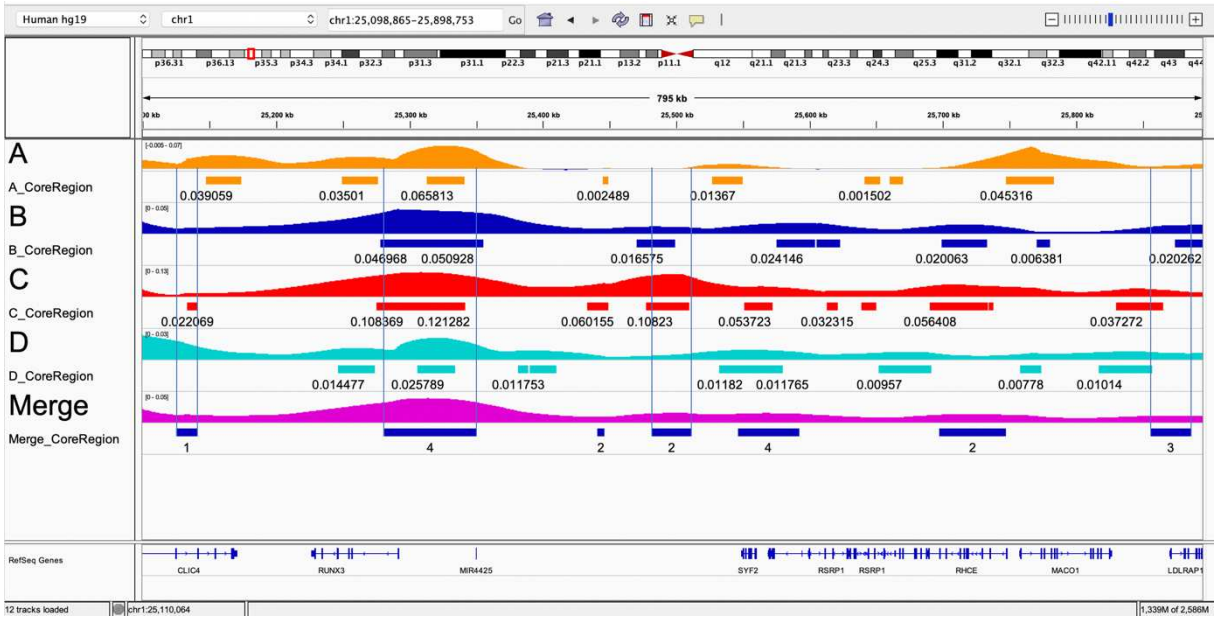


Figure 3.9. The normalized ORM signal distribution and core regions identified in 4 replicates of 0 min data and merged dataset. The values below each core region of individual replicate give the corresponding normalized signal density values, and the numbers below Merge_CoreRegion indicate the overlapped replicates numbers for the corresponding initiation zone core regions. For example, the far-left Merge_CoreRegion only overlaps with red C_CoreRegion within the blue vertical line marked core region in the merged dataset.

Algorithmically, in order to be able to implement such a screening strategy, creating a one-character vector with the total bin number along the genome to record which bins are selected

inside the core regions. The selected ones will be marked as their sample name in lower case such as “a”, “b”, “c”, “d” for 4 replicates, and the outsides of the bin will be labeled as empty string value “”. Then extract these columns in four biological replicates and merge the string items in bins sharing the same genomic position into 1 string value. In this way, I got the new string vector as indicated in Fig. 3.10 for the merged 0 min dataset. The string length of items represents the number of replicates supporting this bin as a core region. For example, “ab”, “bc” with 2 characters means the replicates number is 2, “” empty bin is 0. The content of the merged string tells which replicate datasets they come from.

```
  [1] ""     ""     "a"    "a"    "abc"  "abc"  "abc"  "abc"  "abc"  "abc"  "bc"   "bc"   "bc"   "bc"   "bc"   "bc"   "bc"   "bc"   "bc"   "bc"   "c"    ""     ""     ""     ""     ""
 [27] ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     "b"    "bc"   "bc"   "bc"   "bc"   "bc"   ""     ""     ""     ""     ""     ""     ""     ""
 [53] ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     "bc"   "bcd"  "bcd"  "abcd" "abd"  "a"    ""     ""     ""     ""     "c"    "c"    "c"    ""     ""     ""
 [79] ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     "d"    "d"    "d"    "d"    "ad"   "acd"  "acd"  "acd"  "acd"  "acd"  "acd"  "abcd"
[105] "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abc"  "abc"  "abc"  "abc"  "abc"  "c"    "c"    "c"    "c"    "c"    "c"    "c"    "c"    "abc"  "bc"
[131] "c"    "cd"   "cd"   "cd"   "d"    "d"    "d"    "d"    "d"    "d"    "d"    "d"    "d"    "d"    "ad"   "ad"   "a"    "a"    "a"    "a"    "a"    "a"    "a"    "a"    "a"    "ab"
[157] "ab"   "abc"  "abc"  "abc"  "abc"  "ac"   "ac"   "a"    "a"    "a"    "a"    "a"    "a"    "a"    ""     ""     ""     ""     ""     ""     "c"    "cd"   "bcd"  "bcd"  "bcd"  "bcd"
[183] "bcd"  "bcd"  "bcd"  "bcd"  "bc"   "c"    "a"    ""     "bd"   "bd"   "bd"   "b"    ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     "d"    "bcd"
[209] "bcd"  "bcd"  "bd"   "bd"   "abd"  "abd"  "abcd" "abc"  "ac"   "ac"   "a"    "a"    "a"    "a"    "a"    "a"    "a"    "a"    "a"    "a"    "a"    "a"    "a"    "a"    "a"    "a"
[235] "a"    "a"    "a"    "a"    "c"    "cd"   "cd"   "bcd"  "bcd"  "bcd"  "bc"   "bc"   "bc"   "bc"   "c"    "c"    "c"    "c"    ""     ""     ""     ""     ""     ""     ""     ""
[261] ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     "d"    "cd"   "cd"   "cd"   "cd"   "cd"   "cd"   "cd"   "c"
[287] "c"    "c"    "c"    "c"    "c"    "c"    "ac"   "ac"   "ac"   "ac"   "ac"   "c"    "c"    "c"    "c"    "c"    "c"    "c"    "bc"   "bc"   "bc"   "bc"   "bc"   ""     ""     "d"
[313] "d"    "d"    "d"    ""     ""     ""     "a"    "a"    "abc"  "abc"  "abc"  "b"    ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     "d"    "d"
[339] "cd"   "bcd"  "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abcd" "abd"  "abd"  "a"    "a"    "a"    "a"
[365] ""     ""     ""     ""     "bcd"  "c"    "c"    ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""     ""
[391] ""     ""     ""     ""     ""     ""     ""     ""     "b"    "b"    "b"    "b"    "b"    "bc"   "bc"   "bc"   "bc"   "bc"   "bc"   "bc"   "bc"   "bc"   "bc"   "abc"  "abcd" "abcd" "abcd"
[417] "abcd" "ad"   "ad"   "ad"   "ad"   "ad"   "ad"   "ad"   "ad"   "a"    "a"    "a"    ""     ""     ""     ""     ""     ""     ""     "c"    "c"    "c"    "c"    "c"    "c"    "c"
```

Figure 3.9 Samples lines show the merged adding a column of initiation zone calling output. The string item is corresponding with 1 kb non-overlapping bins along the genome, no matter one given bin whether overlapped with any replicate or not, it will be given a string item. The string item will be “” when there is no replicate overlapped with a given bin. Otherwise, the letters a, b, c, d represent the corresponding replicates: A, B, C, D, overlapped replicas will be added to the string item. And the length of string item will be the number of overlapped replicates.

Then change this merged character vector to an integer vector X1 based on the length of each character item. The X1 contains values 0~4, five different values. For the merged dataset, we also add one column organized by 1/0 as in Figure 3.7 to record which bins are inside core regions. Take this column as a vector X2 with the same length as X1. The bins outside core regions will become elements equal to 0 in X2, and the core region bins correspond to the element equal to 1. Because 0 multiplies any non-zero value is still 0; 1 multiplies any number is still equal to that value. Calculate X3 = X1 * X2, the result X3 will change all non-zero items (1~4) outside the merged dataset core regions into 0. Meanwhile keeps the original value of X1 where the bin located in merged data core regions. Thus, later, according to whether the items in X3 are bigger than 2, we can easily choose the bins inside the merged dataset core regions and with at least 3 replicates supporting these bins in corresponding replicate samples.

3.5.2 The other standard to estimate the quality of core region

3.5.2.1 Normalized signal density is lowly related with core region quality

Besides overlapped replicate numbers, we suppose that normalized signal density is another parameter to estimate the quality of the core region. In Fig 2.18, we found a very good consistency between ORM signal segment enrichment and the RFD curves of OK-seq. For testing, if we can choose a high-quality core region based on the normalized signal density, we introduced the

comparison with the OK-seq initial zones. All core regions were classified into 2 groups depending on whether they were overlapped with OK-seq initial zones or not. The core regions overlapped with the OK-seq initial zones will be considered as highly qualified core regions. Then each group will be further classified into 5 subgroups by the number supporting biological replicates from 0 to 4. The percentage histogram (Fig. 3.10) and histogram (Fig. 3.11) were generated based on the normalized signal density of each initiation zones.

Figure 3.10. Percentage normalized signal density histogram. The x value is the peak point normalized signal density of core regions, and the Y value is the percentage of each subgroup in the corresponding X value (normalize peak density). Different colors represent core regions whether overlapped with OK-seq initial zones and how many supporting replicates show up in the merged core region (from 0 to 4). For the corresponding bins of each X value, the vertical length of the rectangular color block represents the percentage of each category of normalized signal density within the range of the X-axis corresponding to the bin width to which it belongs. R0～R4 means how many replicates support the core region in merged data, the number is 0~4. The categories and corresponding colors in the right legend from top to bottom are like below.

1. ROOverlap (Light red): R0 overlapped with OK-seq initial zone
2. ROUnmap (Brown): R0 not overlapped with OK-seq initial zone
3. Dark green: R1 overlapped with OK-seq initial zone
4. Green: R1 not overlapped with OK-seq initial zone
5. Light green: R2 overlapped with OK-seq initial zone
6. Indigo: R2 not overlapped with OK-seq initial zone
7. Blue: R3 overlapped with OK-seq initial zone
8. Dark purple: R3 not overlapped with OK-seq initial zone
9. Light purple: R4 overlapped with OK-seq initial zone
10. Red: R4 not overlapped with OK-seq initial zone

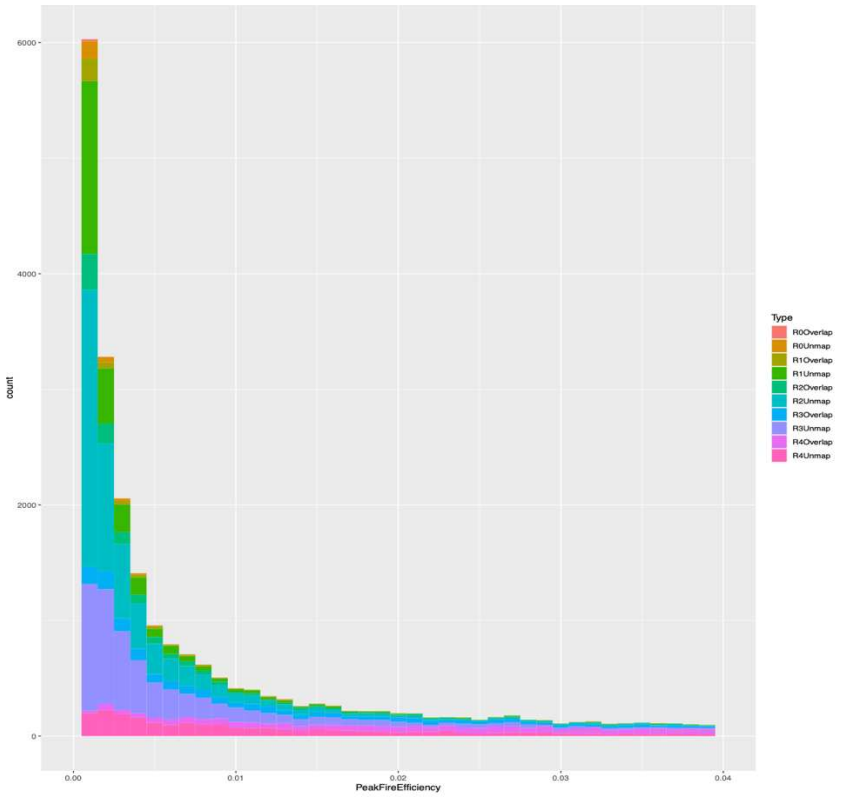


Figure 3.11. Normalized signal density histogram. The x value is the peak point normalized signal density of the core region, and the Y value is the core region number of each subgroup superimpose in corresponding X value. The color code is the same as Fig. 3.10.

The percentage normalized signal density histogram (Figure 3.10) shows that the core regions being detected by more replicates (e.g. 3 or 4) increases in function of firing efficiency, while the percentages of core regions overlapped with OK-seq initiation zones are very low for all groups and does not increase in function of firing efficiency. And based on the normalized signal density histogram (Figure 3.11), we know the total core region number decrease a lot within the range 0~0.02. So, for a very limited number of core regions with normalized signal density bigger than 0.02. Such a low percentage of overlapped core region suggests that the absolute normalized ORM density (associated with firing efficiency) might introduce higher false positive detection, thus ORM signal density is not a good parameter used to perform correct filtering in selecting the initiation zones.

3.5.2.2 ORM signal amplitude and core region quality

In order to balance quality and quantity, we have found a better classification parameter: the ORM signal amplitude, i.e. the difference between the ORM signal density of the peak position and the valley of each peak area (or called Delta ORM value). And the reason why the normalized signal density can't distinguish the high-quality core regions directly. Because what really determines the quality of the core region is the steepness of the shape of the peak, and it does not necessarily depend entirely on the height of peaks.

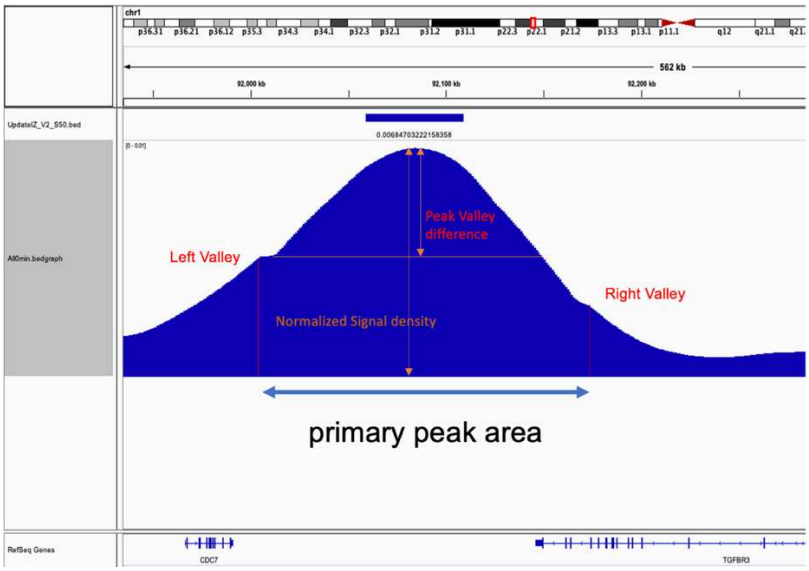


Figure 3.12. An example in calculation of ORM signal amplitude of a primary peak area. The range from the peak point to 2 sides closet valleys is the primary peak area. We take the primary peak area as a unit and choose the valley with a higher normalized signal density to calculate the difference between the normalized signal density of the peak point to that of the valley. This value reflects the steepness of the peak.

With the help of ORM signal amplitude, we can filter the noise peaks with high normalized ORM signal density, but the shape is more or less flat as shown in Fig. 3.13.

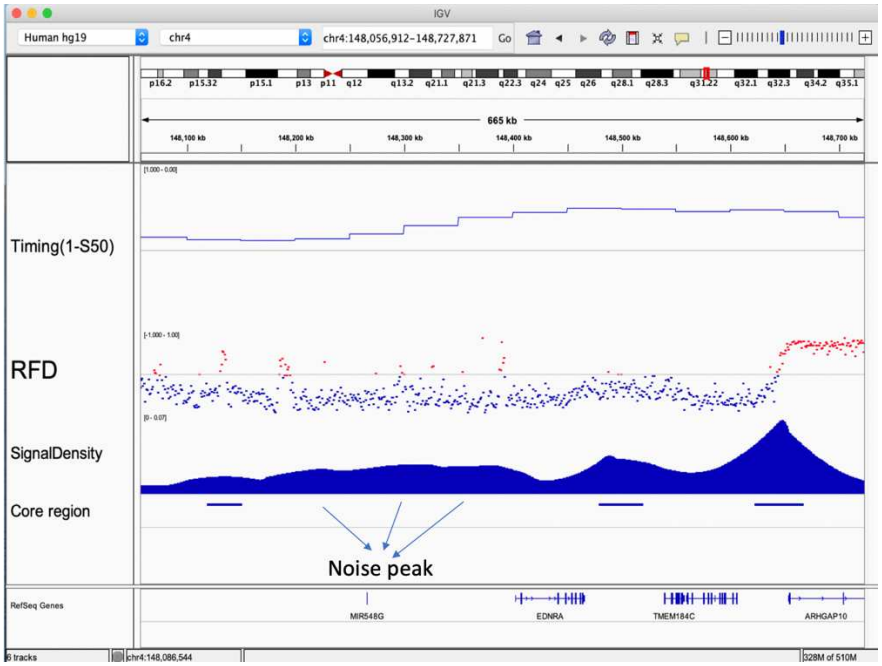


Figure 3.13. Using ORM signal amplitude to filter out the noise peaks. The first line represents the replication timing, it is the 1-S50 value, and the higher it is the earlier it replicated. The Second line is the RFD curves of OK-seq (as in Fig 1.13), and the third line is the normalized signal density distribution. The rectangle bars in the fourth line are the core regions that pass the ORM signal amplitude filtering, and the arrows point to some filtered noise peaks with higher ORM signal density but lower ORM signal amplitude values.

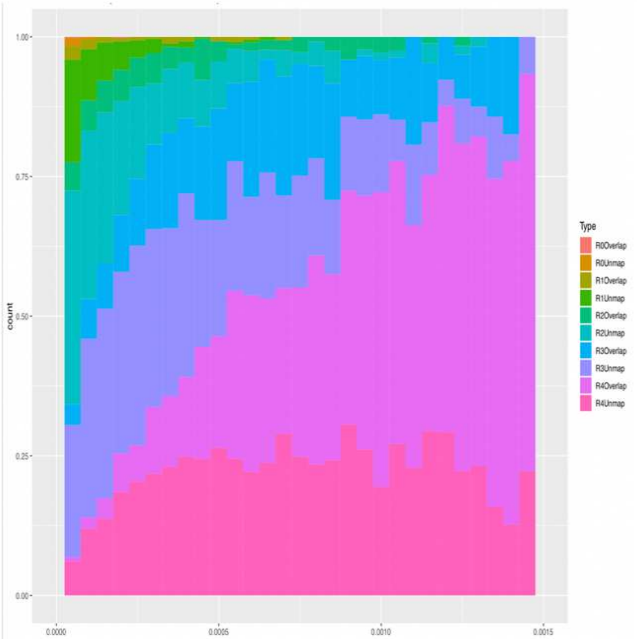


Figure 3.14. Percentage histogram comparison in the function of ORM signal amplitude and normalized signal density. Color labels are similar to Fig. 3.10 but with ORM signal amplitude instead of ORM signal density on the X-axis.

As shown in Fig. 3.14, both (i) the percentage of the core regions being detected by more replicates (e.g. 3 or 4) and (ii) and the percentages of core regions overlapped with OK-seq increases in function of ORM signal amplitude. This highly suggests that comparing to normalized signal density, peak valley difference has a stronger ability to distinguish and use for selecting the high-quality core regions.

Finally, we selected 4,930 initial zones from core regions by combing ORM signal amplitude and replicate number filtering. Only core regions identified in at least 3 replicates and with a relatively steep peak shape in the combined dataset (with ORM signal amplitude greater than 0.3%) were retained as final initiation zones.

3.5.3 K-means clustering for IZ length adjustment

When we got the 4,930 initial zones, we calculated the average distribution of fire efficiency, RFD value, and replication timing (1-S50) around the center of all initial zones to test the quality of the initial zones (Fig. 3.15). And each of them performs well as we expected. Especially, the RFD curve, shows a beautiful S shape, meaning that the initiation zones identified by ORM data are also supported by the OK-seq data.

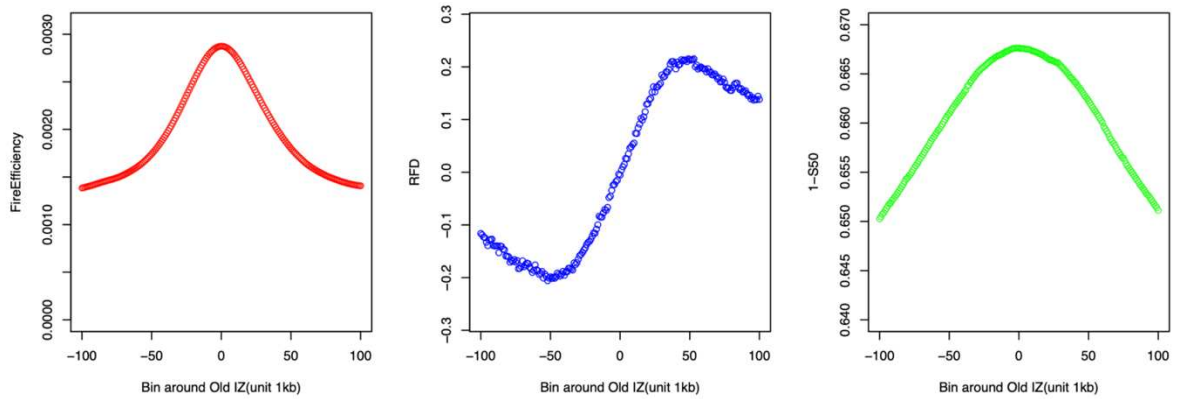


Figure 3.15. The average distribution of fire efficiency (red), RFD values of OK-seq (blue), and Replication Timing values (1-S50, green) around ORM IZ centers. The peak of fire efficiency (i.e. normalized ORM density profile) and Replication Timing show that the initial zone centers are replicated relatively earlier comparing with the neighboring regions. A positive transition observed in the RFD curve is evidence supported that they are bone fine replication initiation zones as illustrated in Fig. 1.12.

As shown in Fig 3.15, the center positions of initial zones identified from the ORM data are reliable. However, the average length of the ORM initial zone is up to 46.9 kb (Fig. 3.18), which is a little bit longer than OK-seq initiation zones (31.2 kb). This may be due to the conservative, a larger percentage 40% cutoff used in the ORM initiation zone calling, resulted in the longer initial zones. All of this may reveal that the initial zones might have an inappropriate length although at the correct genomic position. Therefore, we decided to redo a better estimation of the initiation zone boundaries at the identified initial zone positions. In order to adapt to the different signal distribution of the individual initial zone, we no longer use a uniform signal percentage cutoff but find the appropriate length according to its local distribution shape by k-means clustering.

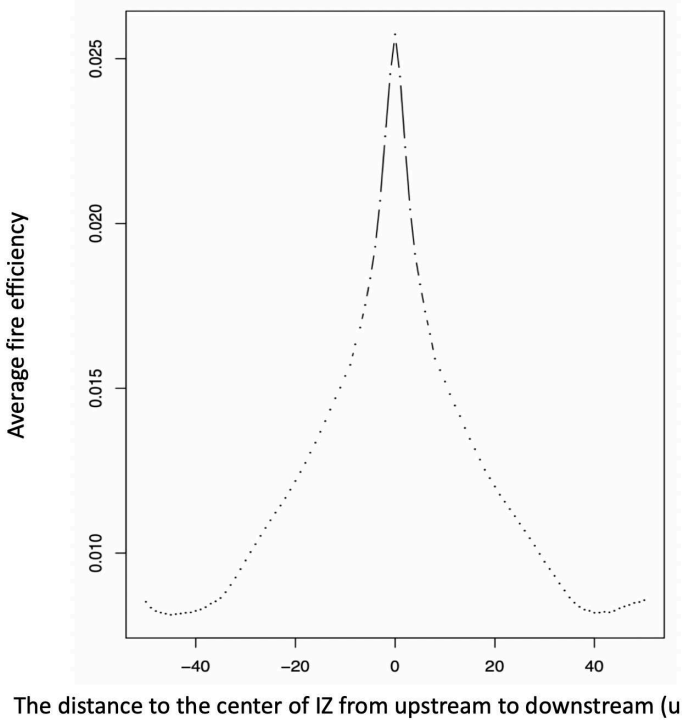


Figure 3.16. The average fire efficiency distribution around the centers of ORM IZs (Initiation Zones).

In Fig. 3.16, we can see the fire efficiency decreases very quickly within +/- 10 kb. The regions range from 10 kb to 30 kb (-30 ~ -10 and 10~30), and the decreasing speed is relatively stable. When the distance up to more than 30kb away from centers, the decreasing speed will become further slower.

Based on this, I took the absolute values of the difference in fire efficiency between all adjacent bins for the primary peak areas of initiation zones (IZ), which is used to calculate the speed of the fire efficiency change (i.e. Delta ORM signal density between adjacent bins). For these differences, I first performed the 3-core k-means clustering. But I quickly discovered that 3-core clustering is likely to interrupt the continuous central area due to noise data, resulting in the calling initial zone too short, especially for some sharp peaks, the IZs will become very narrow. In extreme cases, there is even only a single 1 kb bin containing the peak position. Then I tried dual-core clustering. This time, it caused the average calling initial zone too large, more than 60 kb. So, selecting an IZ length cutoff to classify speed values into two 2/3 groups is necessary, and finally, by various testing, I decided to cluster the difference values into 2 groups at first. Then, if calling IZ length was bigger than 30 kb (based on Fig. 3.16), I further clustered them into 3 groups by K-means (Fig. 3.17).

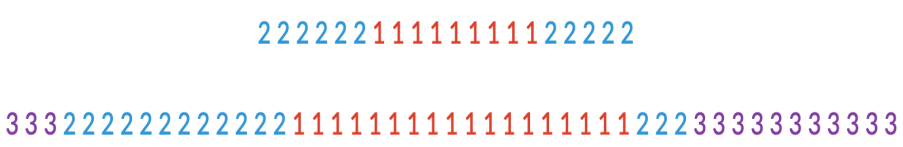


Figure 3.17. Schematic diagram of k-means clustering results.

Firstly, calling IZs must contain peak position, and the IZs maybe not symmetry with the center of the peak based on the native background shape. The new method will check the absolute difference of peak-left-bin difference and peak-right-bin difference. If they are in the same cluster, only need to extend the bin until meeting the first bin belong to another cluster. However, sometimes when the shape like a cliff, the two sides adjacent bin of a peak belong to different clusters. This time, I took the smaller difference value because its fire efficiency is higher (closer to the fire efficiency of the corresponding peak) and stretch to one side until meeting the bin belongs to another cluster.

After applying the k-means clustering calling method to the previous 40% calling IZs, the average IZ length decreased to 38 kb (Fig. 3.18) and I also got a better agreement with the OK-seq and Replication Timing data (Fig. 3.19). The detailed comparison of ORM IZs with replication origins identified by other methods, as well as the genetic and epigenetic features are further described in Chapter 5.

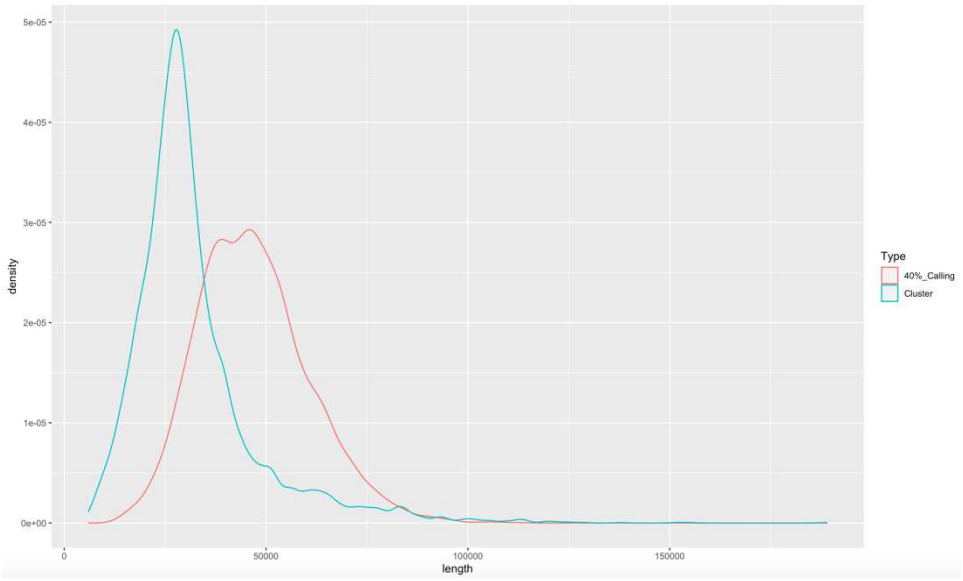


Figure 3.18. the IZ length distribution comparison between 40% signal percentage calling and k-means cluster.

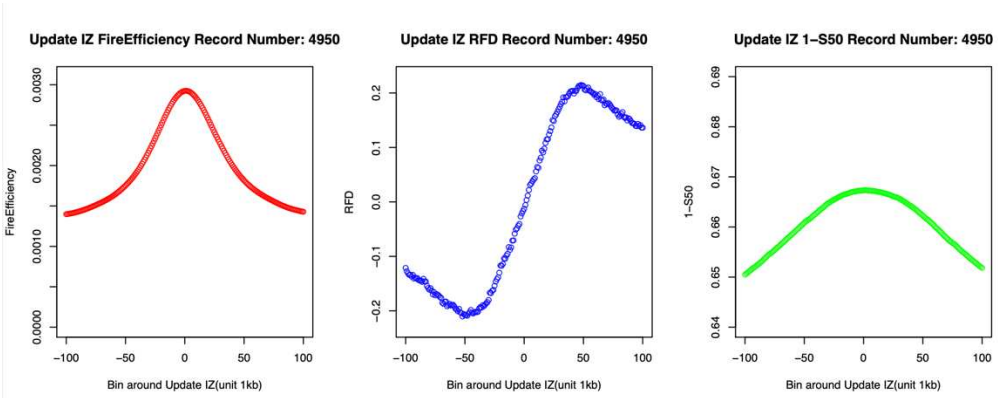


Figure 3.19. The average distribution of fire efficiency (red), RFD values of OK-seq (blue), and Replication Timing values (1-S50, green) around the final ORM IZ centers. The same as figure 3.15 but with the final ORM initiation zones defined by the k-means method.

CHAPTER 4

Fork directionality analysis

4.1 FDI: Fork direction index

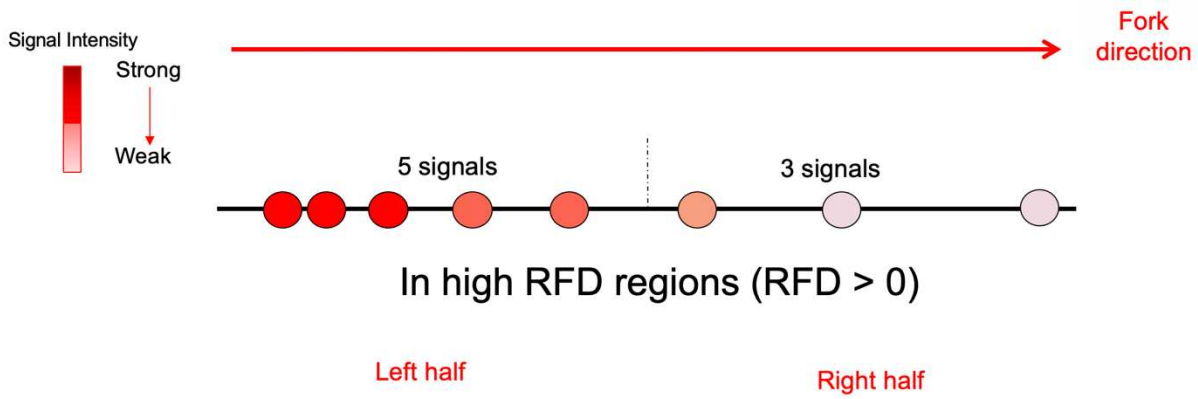


Figure 4.1. Schematic of ORM red signal intensity and density change on an ORM segment representing a right-ward replication fork. The color from dark red to pink represents the signal intensity from strong to weak, the signal intensity sum was calculated at the left and right half of the ORM segment, respectively. Both signal intensity and density decrease along the ongoing direction of the replication fork.

Because the labeling dying is limited, the signal density (and maybe also intensity) will decrease in the orientation where replication forks moving. In addition, at the beginning of labeling, when the signal density is enough high, the camera of the Bionano platform (current resolution ~1 kb) might not be able to distinguish some very closed signals so that would recognize these signal cluster as one single signal with very strong intensity. In this way, the signals on ORM tracks show a signal polarity, which can indicate the direction of replication fork movement. We defined a value, called FDI (Fork Direction Index), calculated by following formula (4.1) according to this property to classify the direction of ORM segments. When FDI > 0, it represents a rightward replication fork, and FDI<0 represents a leftward replication fork.

$$FDI = \log2(IntensitySUM\_Left / IntensitySUM\_Right) \qquad (4.1)$$

For testing the signal strength analysis (Fig. 4.1), we picked out 617 high RFD regions from OK-seq data (Petryk et al., 2016). Within these High RFD regions, >80% reads are forks moving towards one direction (either left or right, based on the average RFD values) (Fig. 4.2), the ORM segments in these regions should show corresponding signal polarity.

For getting more separated replication forks in high RFD regions, we can't use the datasets too early or late during the S phase. At the very beginning of the S phase (e.g. D.0 and D.5 in Table 2.1), most of the synchronized cells just form the replication origins, and in too late of S phase (e.g. D.60 and D.90 in Table 2.1), most of the cells may have finished the replication in high RFD regions.

So, we choose the D.30 dataset (Table 2.1), which starts to label in 30 min after entry of the S phase. For making the signal polarity more obvious, we further designed an experiment, in which after entry of S phase in 30 min we transfected fluorescent dUTP to label the on-going replication forks, then we immediately diluted the fluorescent dUTP concentration by artificially adding buffer after 10 min or 20 min of labeling. In this way, we got 3 replicates of 30min datasets: the one without buffer (30min), the one adding buffer after 10 min of labeling (30min-10), and the one adding buffer after 20 min (30min-20) of labeling.

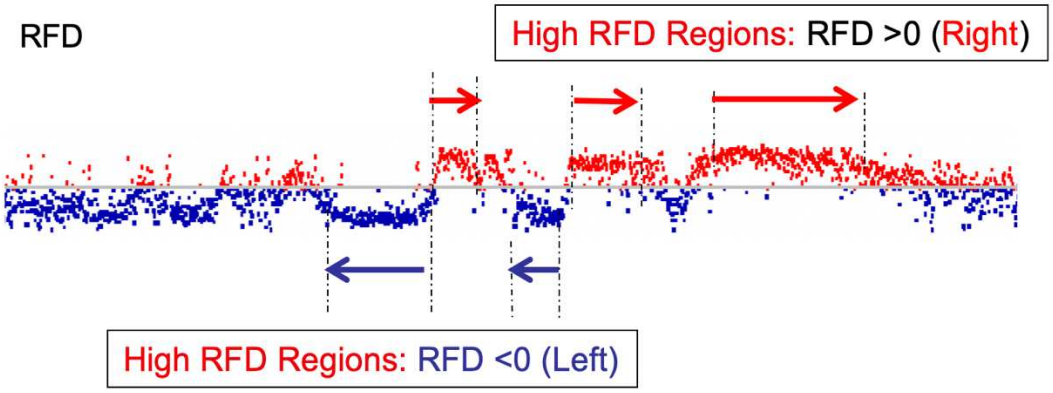


Figure 4. 2. The RFD curve and high RFD regions. The RFD profile obtained by OK-seq as in Fig. 1.13. The dashed line marked the high RFD regions in which more than 80% of Okazaki fragments inside are from replication forks in the same direction. The sign of RFD values represents the shared direction of these replication forks.

I extracted all the ORM segments in these 3 replicates of 30 min datasets, which are overlapped with the High RFD regions. I then calculated the FDI distribution of ORM segments within the rightward and leftward high RFD regions, respectively. In all three replicates, the general direction revealed by FDI distribution consists of the sign of RFD (Fig. 4.3). Unfortunately, for unknown reasons, adding buffer at 10 or 20 mins after labeling did not show any difference on the FDI distribution. This might due to the low labeling efficiency as described in previous sections.

To further verify the correctness of FDI in population-based data, we checked the signal polarity around T-peaks for 20 min, 45 min and 90 min datasets. All results showed an obvious FDI medium increasing trend from negative to positive values around T-peaks (Fig. 4.4). In addition, the regions with polarity expand following the movement of replication forks on both sides (Fig. 4.4).

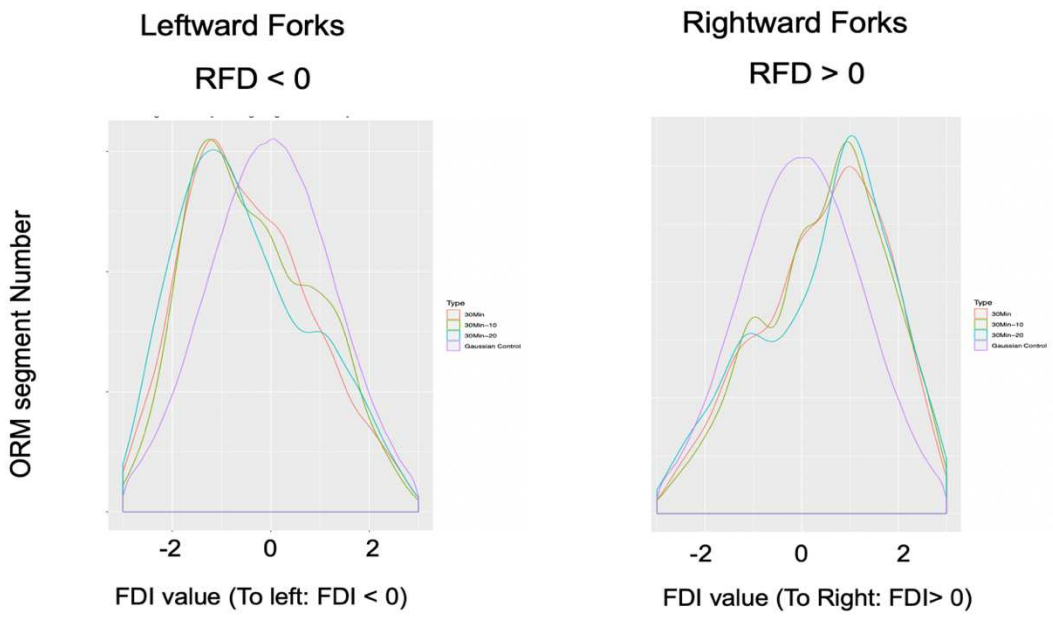


Figure 4. 3. Distribution of Fork Direction Index (FDI) for the ORM segments within leftward or rightward high RFD regions detected by OK-Seq. The purple line is a bell shape Gaussian distribution as the control group, which correspond to the FDI calculated by all ORM segments in 30 min sample. The other three colors are the FDI distributions of ORM segments within High RFD regions in three 30 min replicates, i.e. 30min, 30min-10 and 30 min-20. When RFD < 0, the majority of FDI values are also negative, and *vice versa*.

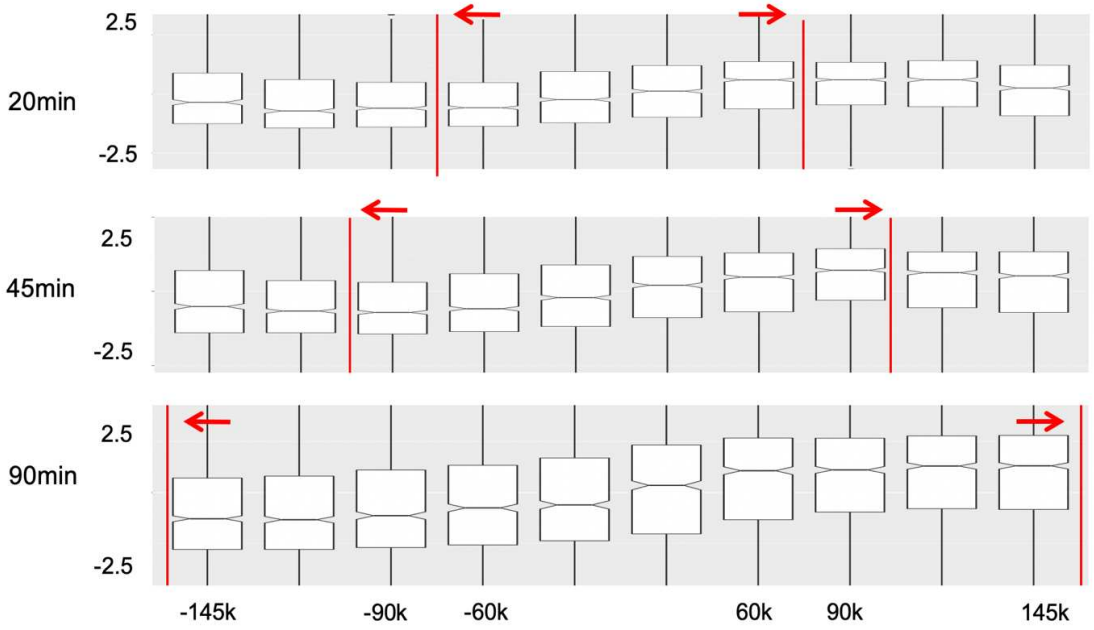


Figure 4.4. FDI distribution of ORM segments around early T-peak centers. The box plot shows the FDI values in each 30 kb windows around T-peak. From downstream to upstream regions, we can observe an obvious change of FDI direction (from negative to positive values) around the center of T-peaks, and the range increases in agree with the fork movement from 20 to 90 min samples.

4.2 The trials for identification of fork direction of individual tracks

4.2.1 The machine learning classifier

In addition to the population average, I tried to use various common machine learning classifiers to further detect the fork direction information on individual ORM segments. First of all, I used 90% ORM segments in all combing 30 min dataset (since the number of ORM segments within high RFD regions is limited) within high RFD regions and corresponding direction labels as training sample. Then I took the last 10% segments as test sample.

The accuracy is up to 60.8% in Support-Vector Machines (SVM), 59.2% in Random Forest and 61.9% in Xgboost (Fig. 4.5), which highly suggests that the fork direction might be predictable for the ORM segments, although at current stage, the error rate is still too high to perform a suitable accurate prediction of fork direction for each individual ORM segment.

Algorithm	Accuracy
SVM	60.8%
RandomForest	59.2%
Xgboost	61.9%

Figure 4. 5. The accuracy of fork direction prediction by 3 different machine learning algorithms. See the main text for detail.

4.2.2 Failed attempt to introduce the second labeling signal

In order to better detect the replication fork polarity on each single replication fork, we tried introducing a second labeling signal to mark the movement of ongoing replication forks. Since the Bionano platform can only use 3 colors at the moment (in our case, 1 for labeling the DNA molecule by blue, 1 for mapping by green and 1 for labeling newly replicated DNA by red), we chose to use green dUTP as the second labeling replicated signals. We assumed that, after mapping the fibers to the reference genome, the unmapped green signals will correspond to the second labeling replicated signals. But to what extend this can be robustly used in ORM detection is not clear, since the introduce of green dUTP as second labeling signals may also lead to mapping error. We designed 2 test experiments. In the first experiment (H9_7030), we used mixed dUTP and second labeling green signals with a concentration ratio of 70% to 30% to mark ongoing replication forks. The idea is that one with higher concentration will be able to label longer period than that with lower concentration. In the second experiment (H9), we used sequential dUTP incorporation

with first red fluo-dUTP and then green fluo-dUTP labeling, in which we expect that the ORM tracks were labeled by 2 kinds of signals one after the other. Unfortunately, based on my preliminary analysis, the red signals and second labeling replicated green signals within the same fiber were far from the expected ratio and order.

In Fig. 4.6, the 2D plot shows the ratio of 2 kinds of labeling signals in first experiment (H9_7030). Each dot means a fiber and the x, y value corresponding to the percentage of second labeling signals (the green signals that are not associated with a mapping site) and red signals in the same fiber. Because we have already known the initial concentration of 2 kinds of signals are 7:3. The black line in the picture is the reference line we draw according to this ratio (7:3). However, almost all the red dots are below this line, and it is obvious that some red dots constitute a straight-line trajectory that meets other ratios, which means the low labeling efficiency of red or both signals.

In Fig 4.7, concerned the second experiment (H9) in which the 2 kinds of labeling signals were incorporated successively. So, the wide and thin sticks which represent 2 kinds of labeling should be more likely to cluster respectively. But the DNA fibers TRPH9_15999 and TRPH9_142580 marked by orange box contain wide sticks mixed with thin sticks, which represents no order in 2 kinds of labeling. The occurrence of this labeling mixture phenomenon has 2 kinds of possibility. One possibility is the second labeling has the effect to the DNA mapping, the wrong mapping may make Bionano take the mapping green signals as second labeling signals. Another is several close origin sites fire together. So, in order to avoid the mutual interference between the green mapping signals and the second labeling green signals, we added a control experiment. In the new control experiment, DLS green, fluorescent labeling for DNA mapping is not used, and only the second labeling replication signals and red signals are retained. The result is much unexpected, because there are only very few fibers containing both red signals second green labeling signals, which means the sequential labeling is not feasible at current stage due to the poor labeling efficiency. This could be due to hard to get 2 kinds of fluorescent labeling sginals, transferred into cell by Electroporation and Nucleofector, onto the same on-going replication forks. This method is difficult to effectively mark ongoing replication regions, especially for second labeling dye.

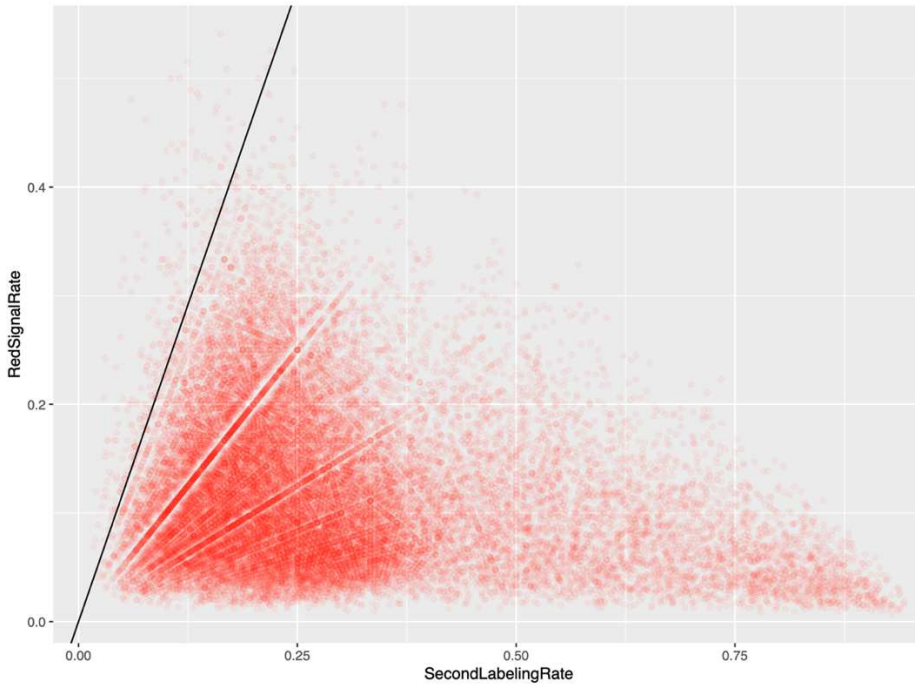


Figure 4. 6. 2D plot for signal density concentration. X axis for the second labeling signal rate per fiber and Y value for the red signal rate. The slope value of black line is 7:3.

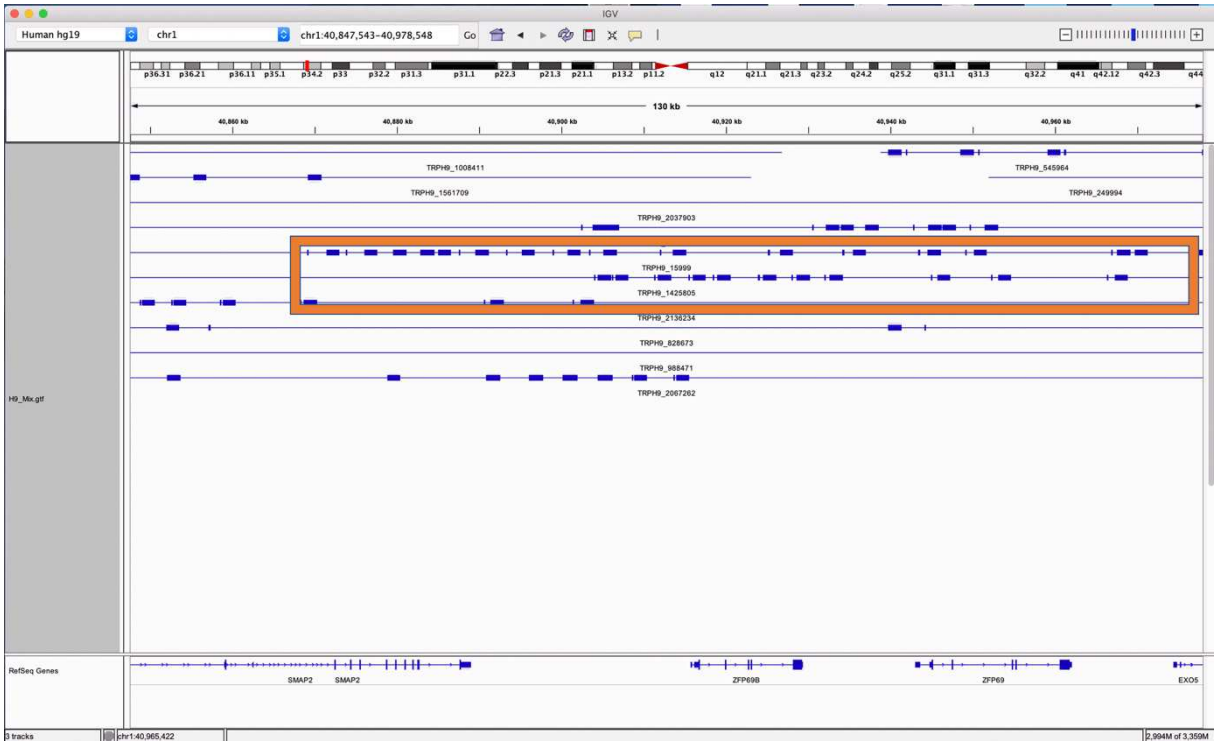


Figure 4.7. The second labeling signal and red signal distribution along the fiber. Each line is a DNA fiber. The wider sticks on DNA fiber are second green labeling signals, and the thin sticks on the DNA fiber are red labeling signals. The orange box marked 2 DNA fibers with abnormal mixed labeling in sequential labeling experiment.

All of above might result from the fact that the mapping of fibers has been affected by the second labeling replicated green signals, and/or only few replication forks have been labeled with both red and green fluo-dUTP due to the low labeling efficiency. Additional works and new experiments need to be performed in the future to further explore the data.

4.3 Genome-Wide Replication Kinetics in Asynchronous Cells

Although it is currently impossible to accurately determine the direction of individual replication fork, as long as there is a sufficient amount of data, FDI obtained from ORM segments can still be used to determine the replication kinetics along the genome by bulk data, like RFD curve in OK-seq. Concerned with most cell types are not amenable to precise cell-cycle synchronization, we decided to use asynchronous data to test the universal value of FDI. In theory, we can also cluster ORM tracks and calculate the FDI value for each track with at least 3 signals in ORM data obtained from asynchronous cells. The sign of FDI reveals the signal polarity of ORM tracks, we took the tracks with +/- FDI values as replication forks to left or right, respectively. Then, for any genomic regions or adjacent bins along the genome, similar as the Okazaki Watson and Okazaki Crick ratio used in OK-seq (Fig. 1.12), we can also use the leftward and rightward fork number within each bin to calculate an RFD value. We call it FDI_RFD. If we average the profile around all initiation zones, the results obtained with FDI_RFD will be totally similar as RFD obtained with OK-seq, but the experiment cost will be much cheaper. Based on above method, we identified 412,113 replication tracks in two biologically-independent HeLa replicates totaling 1.4 Tb of data and 299,595 tracks in one H9 dataset totaling 738 Gb of data (Table 2.1). Using the same analysis as for our synchronous datasets (section 2.8.4), we inferred a similar nucleotide-labeling frequency of 1/1025 and 1/850 thymidines, respectively. The replication tracks average 23.9±35.5 kb in length in the HeLa data and 27.5±40.4 kb in length in the H9 data, which again is comparable to the length of tracks in the synchronized data. Thus, forks released from aphidicolin arrest synthesize at about the same rate as untreated replication forks. The tracks are uniformly distributed across the genome, as predicted for asynchronous replication forks, with an average density of 1.3±0.5%. In particular, in contrast to our synchronous dataset, and as expected, we see no enrichment at replication timing peaks in early- or late-replicating regions (Fig. 4.8).

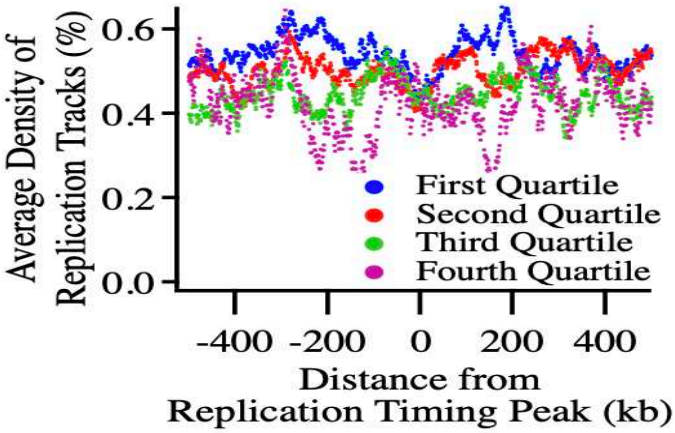


Figure 4. 8. No ORM signal enrichment around replication timing peaks in asynchronous data. Take the timing of T-peak centers, we rank all timing values and classify to 4 groups from early to late by quartile (Early: first blue quantile, MidEarly: second red quantile, Midlate: third green quantile, Late: fourth purple quantile), then aligned all track center to 1 kb bins along the genome, as in Figure 2.1 to normalized track number in each 1 kb bin and get the average density distribution of replication tracks around all replication Timing Peak regions (+/- 500 kb)

Meanwhile, as expected, the FDI_RFD calculated by ORM data performs a similar trend of RFD obtained with OK-seq, no matter along the chromosome (Fig. 4.9) or around the T-peaks (Fig. 4.10).

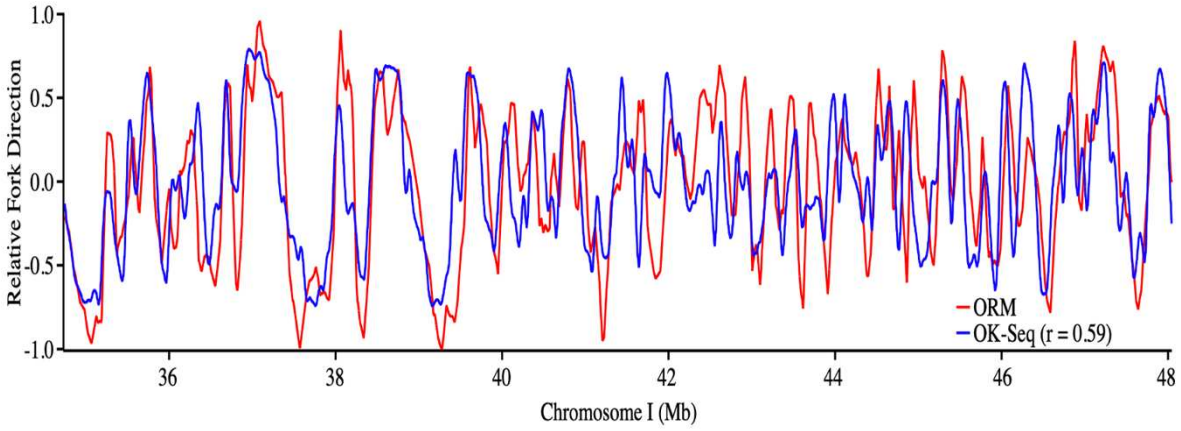


Figure 4. 9. The comparison of FDI_RFD curves of ORM and RFD curves of OK-seq along chromosome 1. Two kinds of RFD curves calculated by different methods, the red profile is FDI_RFD of ORM, and the blue one is the RFD of OK-seq. The Pearson score between them up to 0.59

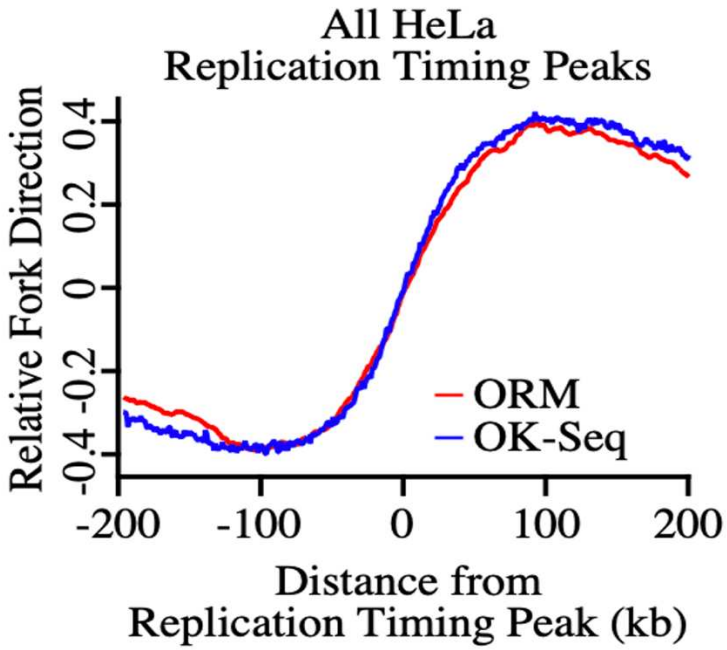


Figure 4. 10. The comparison of average FDI_RFD curves of ORM and average RFD curves of OK-seq around the T-peak regions. Two kinds of RFD curves calculated by different methods within the +/-200 kb range around T-peak regions. The red profile is FDI_RFD of ORM, and the blue one is the RFD of OK-seq. The shape of two kinds of profiles consists of each other.

Importantly, the polarity signal in ORM data is cell-type specific (Fig. 4.11). These results show that ORM data can be used to characterize replication kinetics in unsynchronized cells, demonstrating its applicability to any cells that can be pulse-labeled with fluorescent nucleotides.

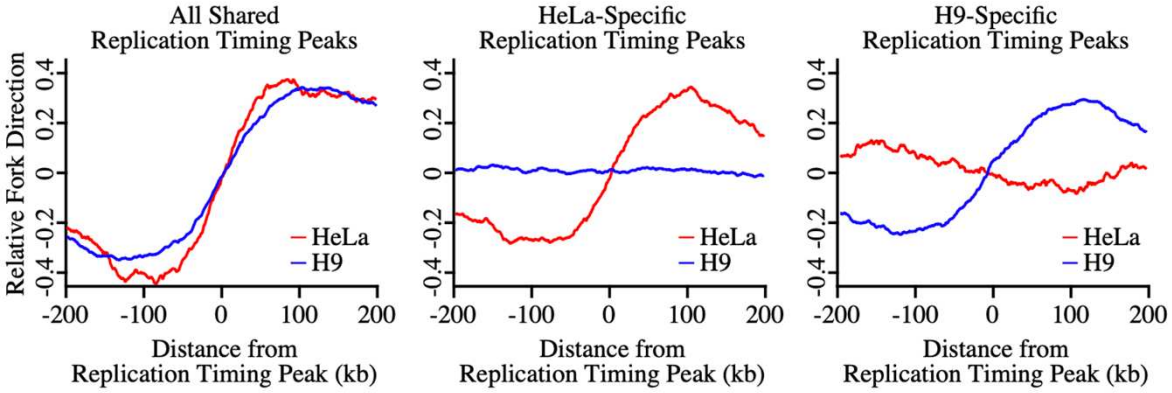


Figure 4. 11. The comparison of average FDI_RFD curves in 3 kinds of T-peak regions between Hela cell line and H9 cell line. The red profile is for Hela and the blue one is for H9. The far-left plot is FDI_RFD curves around shared T-peaks regions in both cell lines. The middle one represents the FDI_RFD profiles in Hela specific T-peak regions. And the right plot is the FDI_RFD profiles in H9 specific T-peak regions. No matter Hela or H9, the FDI_RFD only show S shape in the T-peak regions belong to their own cell line.

CHAPTER 5

Deeper derivative data mining for ORM IZs

5.1 Stochastic model

5.1.1 Early initiation events in late-replicating domains

Because our data is from synchronized cells at the beginning of S phase, theoretically, all detected signals should be from regions with early replication timing based on deterministic model. However, in Figure 5.1, there are still ~9% early replicated ORM track centers located within regions, which are recognized as late replicating domains by Repli-seq (population average replication timing). Thus, our ORM data doesn't meet the deterministic model.

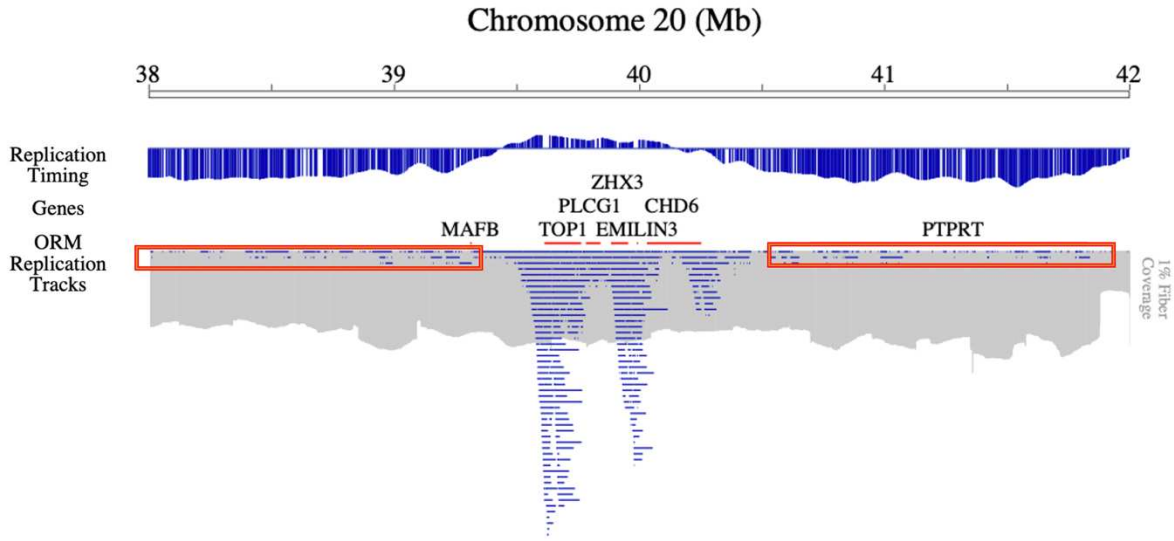


Figure 5.1. ORM tracks distribution along a genomic region on Chromosome 20. The first line shows the Repli-seq timing distribution. It is the average replication timing value based on population-based data. The negative timing in the 2 sides represents late replication domains, and the mid part is an early replication domain. The second line is gene names. The third line shows ORM tracks obtained in the 0 min dataset. As expected, the majority of ORM tracks (91%) enriched in the early domain and associated with active genes, such as TOP1 and EMILIN3. But there are also numerous ORM tracks marked by the red boxes on both sides within the late-replicating domains.

5.1.2 Late-replicating signals are not noise data

Although we detected the ORM tracks from late domains, they might result from technical noise. To verify this hypnosis, here, I grouped T-peak regions into four quartiles based on the S50 timing value in the center of reigns (Chen et al., 2011). The 1st to 4th quantiles is a group containing T-peak region centers with timing from early to late in the order. So, the late replicating domains correspond to the 3rd quartile (S50 between 0.5~0.75) and the 4th quartile (S50 > 0.75). In order to test whether ORM signals are really consistent with the stochastic model, we need to check if the ORM signals can really reveal the initiation process in the 3rd and 4th quartiles.

First of all, I did the segmentation clustering by GMM (Gaussian mixture model) to get the ORM tracks in asynchronous data including the solo track with only one signal. Because in asynchronous data, the ORM tracks correspond majority of ongoing replication forks distributes randomly, and there is no obvious initiation enriched region around IZs observed in the 0 min data, I calculated the ORM signal density around T-peak regions to observe the ORM signal distribution of asynchronous Hela cell line. As expected, there is no enrichment at centers of replication timing peaks in each quartile of ORM tracks (Fig. 4.8).

Then, I came to a synchronized 0' min dataset, in which ORM tracks enrich around replication initiation sites. Similarly, we also group T-peak regions into 4 groups and observe the ORM signals distribution around T-peaks in synchronized cells. In Fig. 5.2, although the signal density level in the 3rd quantile group is much lower than the 1st and 2nd groups, we observed a similar enrichment of 1st, 2nd and 3rd quantiles comparing with the surrounding regions of each group, all of which show an SNR (signal-noise ratio) equal to ~2.8. The average signal density enrichment of the 3rd quantile confirmed that the replication initiation really occurs in late replication domains within some early S phase cells. It is to say that the ORM data may be consistent with a stochastic firing model.

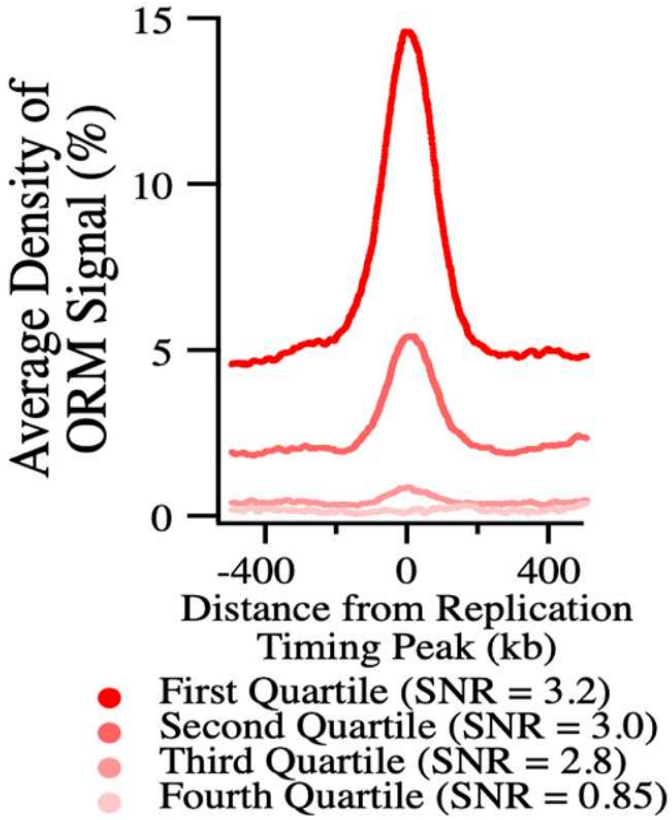


Figure 5. 2. ORM track density around T-peaks in four replication timing groups. The red colors from dark to light represent T-peaks belong to the 1st quantile to 4th quantile defined as in Fig. 5.2. The X-axis shows the distance to the center of T-peak regions from upstream to downstream. The Y-axis gives the average ORM signal density. The signal density is calculated by mapped ORM signal number to divide the mapped fiber number per 1 kb bin.

5.1.3 Firing efficiency is correlated with replication timing

In fact, based on Figure 5.1, it is not hard to see, the replication timing is earlier the fire efficiency (estimated by normalized ORM track density) will be higher. To verify this idea, we calculated the correlation between the fire efficiency of 4,930 ORM initial zones defined by fire efficiency steep peaks and the S50 timing value correspond to the center of each IZ. The Pearson score can be up to 0.75 (Fig. 5.3). It indicates that the fire efficiency of ORM initiation zones is highly correlated with replication timing. Moreover, normalized ORM track density calculated within a given position could be recognized as the probability for initiation occurred in this position.

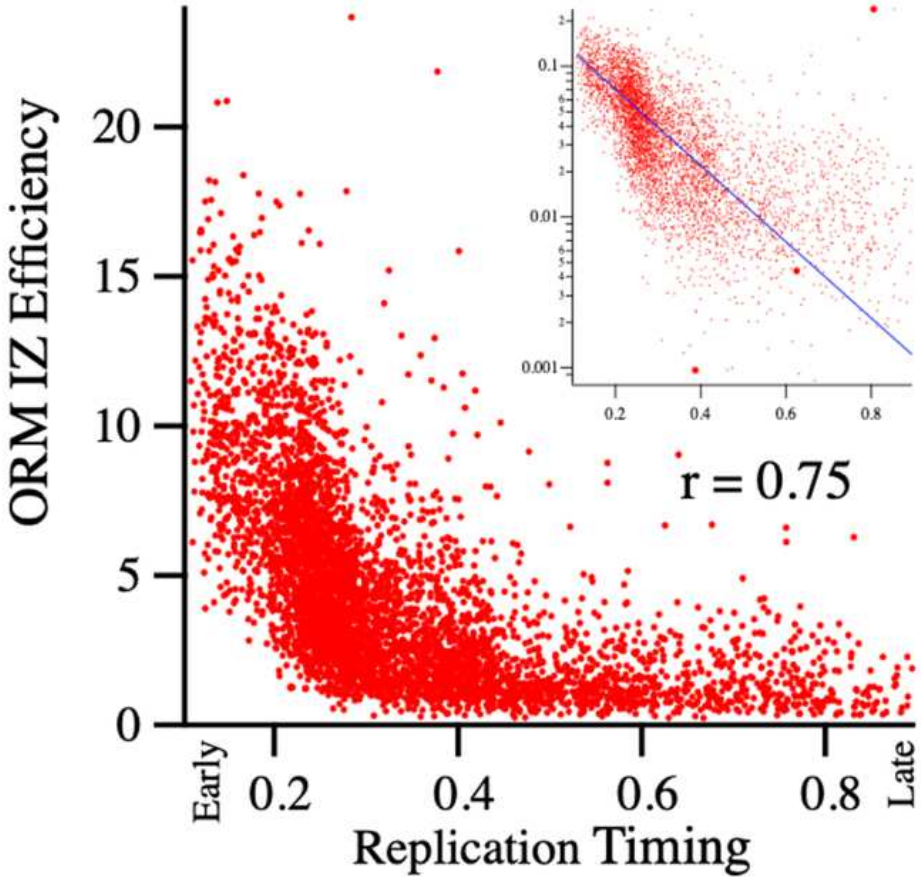


Figure 5. 3. The 2D dots plot for replication timing and fire efficiency of ORM initiation zones. The upper right plot is a thumbnail after compressing the Y value of fire efficiency by log transformation.

5.1.4 No specific initiation sites

In mammalian DNA replication, whether there is a specific and frequently used replication initiation site has always been uncertain. In our research, we didn't find such kinds of origin sites. Instead, there are initiation events that randomly occur in the broad initial zone. Because the resolution of ORM is 15 kb, to test the initiation events come from different origins, we count the number of ORM tracks with 15 kb inter-distance (the distance between two adjacent tracks' centers) inside each IZs (Figure 5.4). We call this number for a given IZ tilling number.

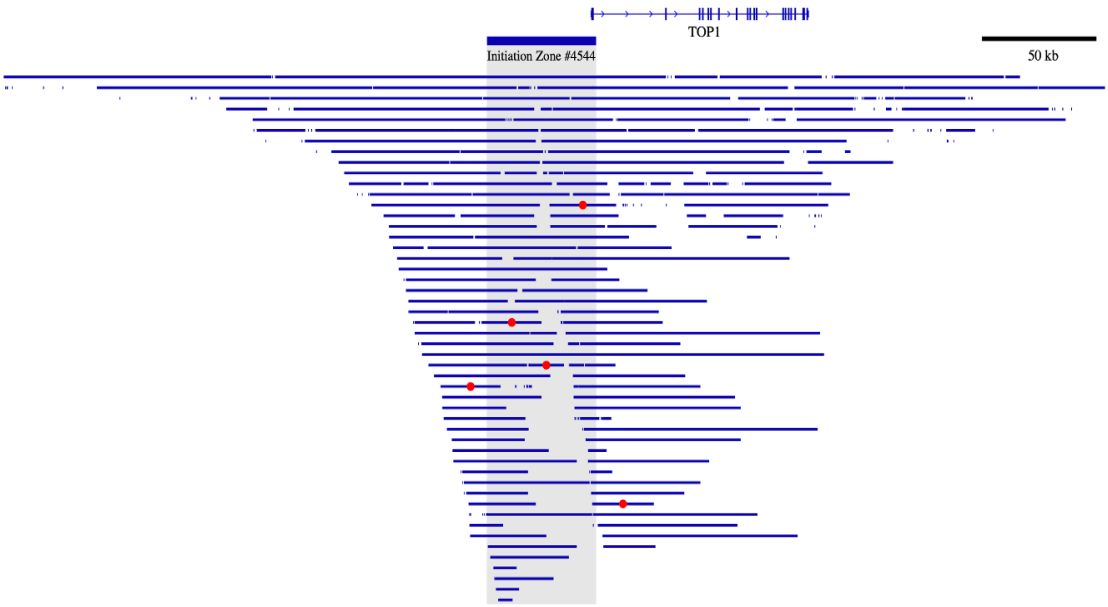


Figure 5.4 The IZs close to gene TOP1 with a tilling number of 5. On the top of the plot is the position of gene TOP1. The second line is the closet IZ. The region of IZ is also marked by the shadow below. Below the IZ, is the enrichment of ORM tracks. The red dot is the center of marked ORM tracks overlapped with shadow regions and the distance between red dots is at least 15 kb. These 5 red dots represent there are at least 5 independent non-overlapping initiation events that occurred around the initiation zones.

Some research reports that gene loci around Top1 are identified as specific replication sites, (Keller et al., 2002; Tao et al., 2000). However, as shown in Fig. 5.4, the tilling number of nearby IZ is 5, which means the initiation event occurred stochastically based on the ORM track's distribution. Not only for this specific IZ, 93.49% IZs contain at least 2 ORM tracks with 15 kb inter-distance (Tilling number >1). Figure 5.5 shows the distribution of the tilling number in all IZs.

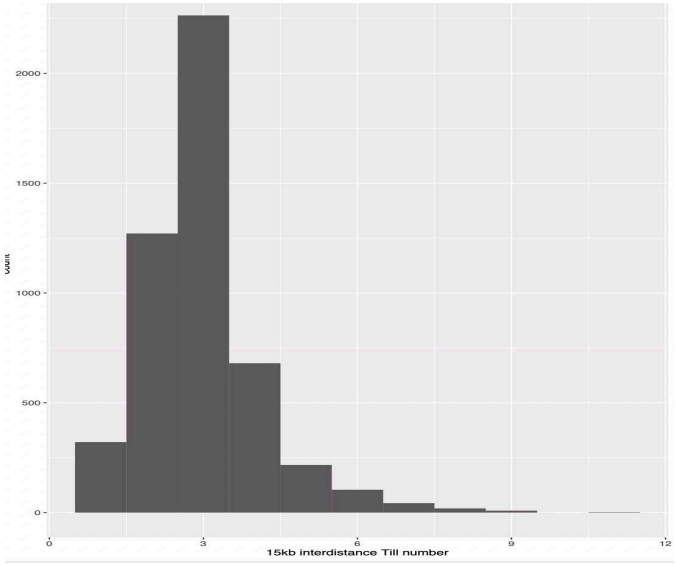


Figure 5.5 The histogram of tilling number for all IZs. The tilling number is the estimated minimum number of initiation sites per IZ. The estimate was made by calculating the minimum number of replication tracks in each IZ whose centers are more than 15 kb apart. Most of the IZs for which all track centers are within 15 kb of each other are small and contain few replications tracks

5.1.5 Computational simulation confirms the stochastic model

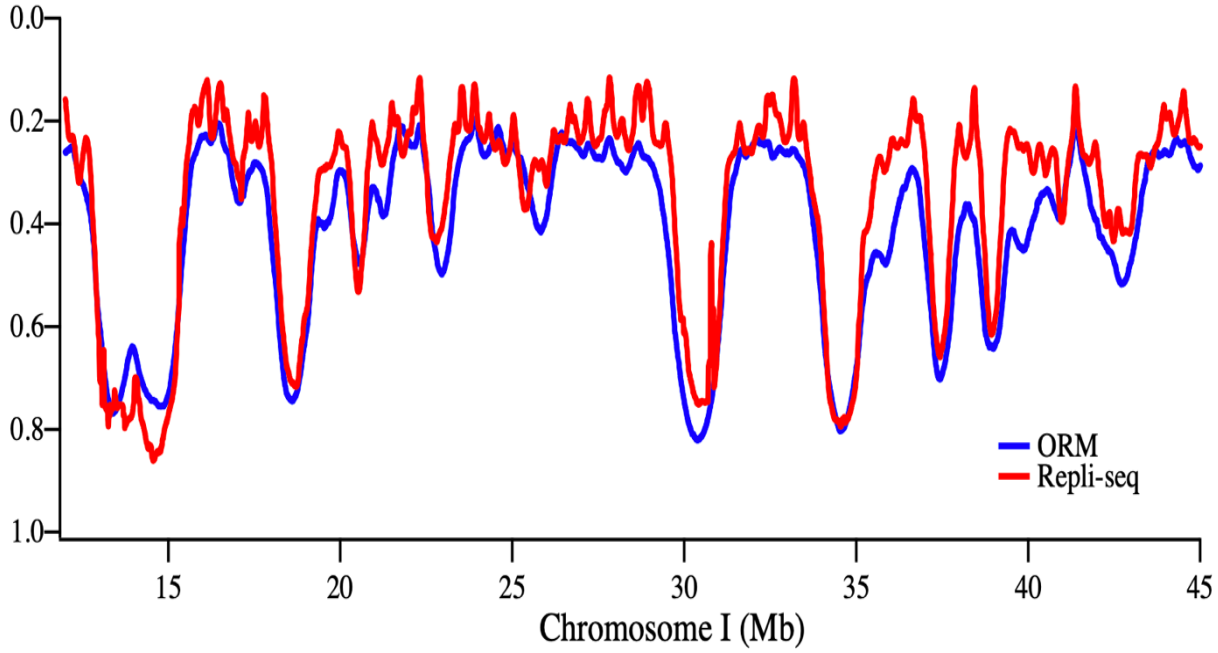


Figure 5. 6. The comparison between experimental and simulated replication timing profiles. The blue and curve is replication timing curves simulated by ORM normalized signal density on chromosome 3 by using the Replicon algorithm (Gindin et al., 2014). And the red line is the experimental replication timing curve S50 (Chen et al., 2010). The Pearson score of simulated replication timing with ORM data and the experimental replication timing data is up to 0.85.

In order to test the stochastic model in timing regulation, I used an analytic model (Yevgeniy Gindin et al, 2014) to simulate the replication process to generate replication timing along the human genome. In this model, the firing efficiency was used to simulate the probability of initiation occurred in a given position: the higher the efficiency is, the bigger probability of the region will be fired at the early S phase. In this way, the simulated replication timing profiles (Fig 5.6) obtained by using origin fire efficiencies calculated from the ORM data consist very well with the experimental replication timing data. This supports the stochastic model of timing regulation.

5.2 Comparison between replication origins mapped by different approaches

5.2.1 Mutual authentication

We further compared the ORM tracks with the replication origins mapped by other methods including OK-seq (Petryk et al., 2016), SNS-seq (Picard et al., 2014), Ini-seq (Langley et al., 2016), and Orc1 ChIP-seq (Dellino et al., 2013), as well as replication timing (Hansen PNAS 2010, Chen Genome Res 2010) in Hela cells. Among them, the origins mapped by OK-seq show the best fitting with ORM segments. We observed significant enrichment of the origins mapped by other methods around the centers of ORM segments (Fig. 5.7A), where OK-Seq showing the highest enrichment, then followed by ORC-ChIP Seq, Ini-Seq, and less extent SNS-Seq. Similar results were observed

by computing the enrichment of ORM segments around the origins mapped by other methods (Fig. 5.7B).

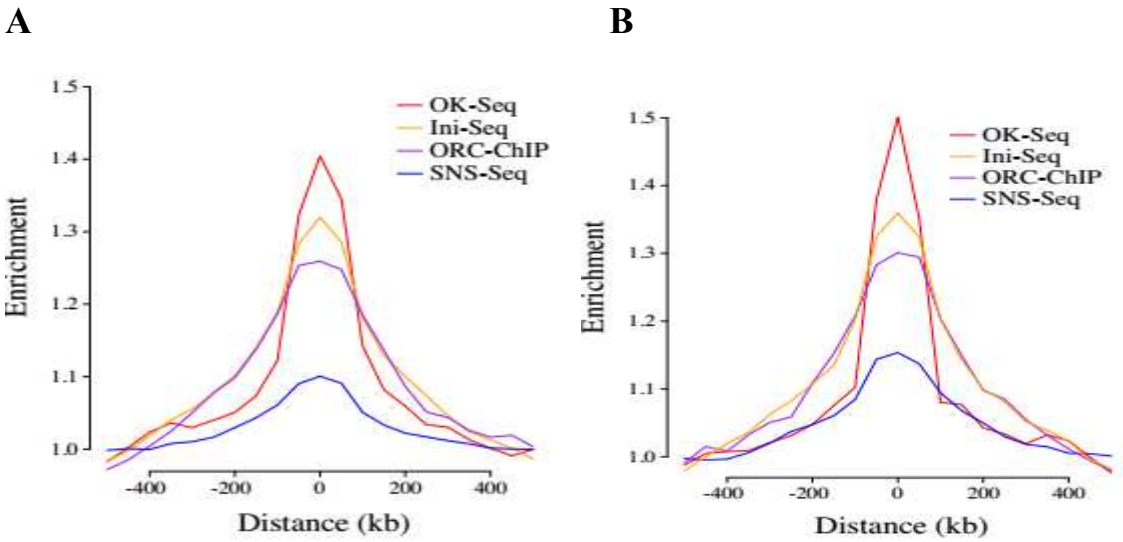


Figure 5. 7. The origin number enriched mutually between ORM tracks and origins identified by other methods. The red, yellow, purple, and blue colors represent 4 kinds of approaches OK-seq, Ini-seq, ORC-Chip, and SNS-seq. We count the number of mapped origin tracks in each 1 kb bin and to see the enrichment of origins identified by different methods around the center of ORM tracks **(A)** and the average mapped ORM track number around the center of origins mapped by different methods **(B)**.

5.2.2 Different fire efficiency and replication timing comparison

Based on figure 5.4, we know the correlation between fire efficiency of initial zones and their replication timing. We compared our ORM data to four published genome-wide HeLa replication-initiation-mapping datasets: OK-seq (Petryk et al., 2016), SNS-seq (Picard et al., 2014), Ini-seq (Langley et al., 2016), and Orc1 ChIP-seq (Dellino et al., 2013). As shown in the previous section, by mutual authentication, replication tracks appear to be enriched around the IZs or origin sites identified by other methods and ORM mutually (Fig. 5.8). And we want to further test the correlation between origin densities of different methods and replication timing along the entire genome. To determine whether the apparent co-localization of initiation mapping data is robust, and to quantify its extent, we measured the correlation between the five datasets (Fig. 5.8). Concerned with the track densities of various methods differ greatly in small interval such as 10 kb along the genome, and the resolution of replication timing is ~100 kb, here, we use 100 kb resolution in our calculation. We found that ORM replication track density correlate well with Ini-seq ($r = 0.59$), to a lesser extent with OK-seq ($r = 0.49$), and followed by SNS-seq ($r = 0.36$) and Orc1 ChIP-seq ($r = 0.31$). These correlations were further confirmed by ROC analysis (Fig. 5.9). The R package involved in the ROC curve drawing is “pROC”. All IZs from different datasets will be applied to a 100 kb adjacent window along the genome. The windows overlapped with ORM IZs will be labeled as “true windows” with an initiation event occurred. Then, we introduced the GLM algorithm and based on if the true windows are overlapped with the IZs from other datasets to calculate the TPR (true positive rate) and FPR (false positive rate) for each approach. According to AUC (area under the ROC curve) values, the correlation of each approach with ORM

is still the same order as the result of the heatmap (Fig.5.7). Ini-seq has highest AUC value (0.73), second is OK-seq (0.72), followed by SNS-seq (0.68) and ORC ChIP-seq (0.62). The replication timing peak as the control group has an AUC value of 0.61.

And in the heatmap (Fig 5.8), we also used the DNase I hypersensitive sites (hereafter DNase I), G4 density, and replication timing values in our analysis. We find a very high correlation between ORM and replication timing (0.82). The DNase I, which represents the open chromatin, also shows a good spearman correlation score (0.73). But the spearman correlation score with G4 is only 0.5 and we can see the correlation score between G4 and replication timing is 0.4, which means the G4 may not really directly related to the replication process.

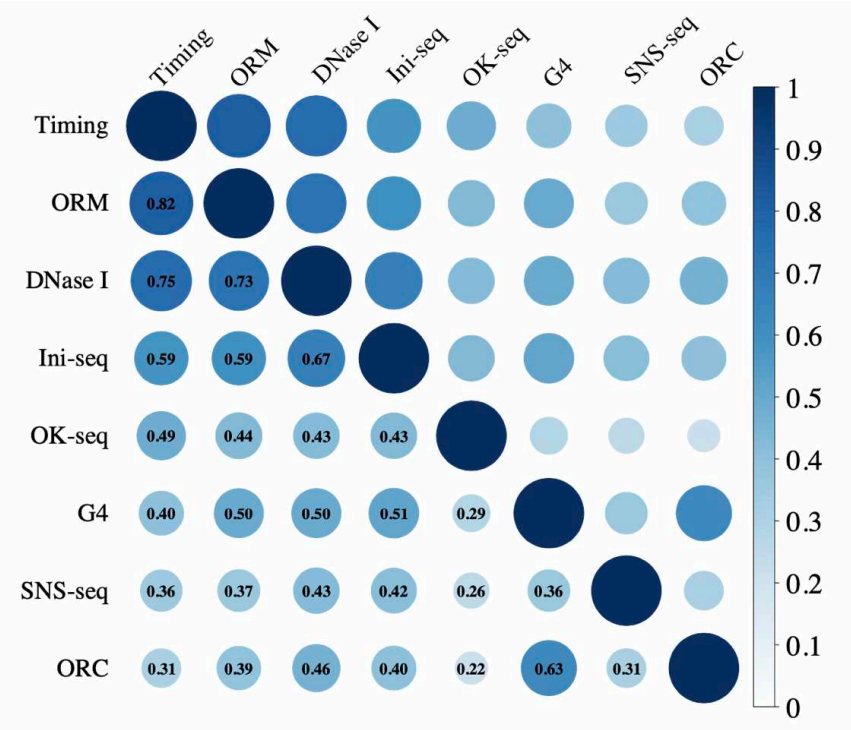


Figure 5. 8. The heatmap shows the spearman scores between origin density mapped by various methods and replication timing, Together with DNase I hypersensitive sites and G4 motifs density. Concerned with the correlation between fire efficiency and timing value is not purely linear, we chose to use Spearman correlation score to test their correlation instead of Pearson correlation score.

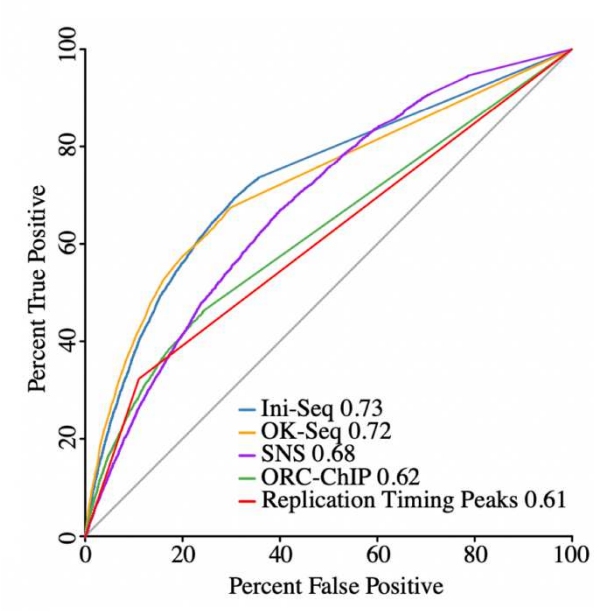


Figure 5. 9. the ROC curves between different origin mapping approaches and replication timing. ROC analysis of the association between the ORM IZs and IZs or origin sites from other methods. As shown in the ROC curves, the ORM IZs are better correlated with Ini-seq IZs (AUC=0.73) and OK-seq IZs (AUC=0.72). The bias of the SNS ROC curve towards high true-positive rates only at high false-negative values is consistent with that dataset having more false-positive signal, whereas the bias of the replication timing and ORC datasets relatively high true-positive rates only at low false-negative values is consistent with those datasets having fewer, but more accurate true positives. Areas under the ROC curves (AUC) are shown in the legend.

5.3 The epigenetic modification marks around initiation zones

Many researchers agreed that replication origins locate in open chromatin regions and might be associated with GC-rich regions (Chevereau et al., 2009; Pope et al., 2011). Another salient feature of origin distribution is its location almost parallels TSSs (transcription start sites), which might suggest a co-evolution of the regulatory regions driving replication and transcription (Sequeira-Mendes et al., 2009). At the same time, to avoid transcription-replication conflicts, the origins should avoid gene coding regions as much as possible, that is why the origins tend to locate in intergenic regions (Liu et al., 2020). There is a recent nature publication also reporting that the H2A.Z facilitates licensing and activation of early replication origins (Long et al., 2020). All above is to say that the occurrence of initiation has significant epigenetic characteristics.

5.3.1 The epigenetic modification marks enriched at ORM initial zones

5.3.1.1 The ChIP-seq data of histone modifications from Encode project

Many genome-wide studies have agreed that the association with the accessible region is especially true in the case of early-replicating origins. However, late-replicating regions are more dispersed and there is no such deterministic connection. So, we still group all IZs into four quantiles based on the replication timing of their center, and then we selected a series of open chromatin histone marks like H3K4me1, H3K9ac, H3K27ac, the histone marks being reported to be related to

replication process like H2A.Z, and the marks related for transcription process, such as H3K79me2, RNA Pol II-phosphoS2. I then tested their distribution around initial zones of four quantile groups. All data in .bw format were downloaded from the Encode project (https://www.encodeproject.org). The detailed downloaded information and related Encode ID for each dataset could be found in the supplementary information.

5.3.1.2 Perform normalization of downloaded .bw files

Considering removing the batch effect, we need to do the normalization for all .bw files and to make sure that all signal strengths are within a similar range of log2 values with the same background. The ChIP-seq strength values are in log2 transformation (log2(ChIP/IP), and I normalized the background level as 0, i.e. the background is the same level as IP. After the normalization, it will be much easier to compare the enrichment situation between different histone modification marks. For realizing that, firstly, I mapped the downloaded ChIP-seq signal strength to 1 kb bins along the genome and calculated the average signal strength within each mapped bin. Then, I drew the average signal strength distribution in a histogram like the example shown in Fig. 5.10 to normalize the average of all samples to 0.

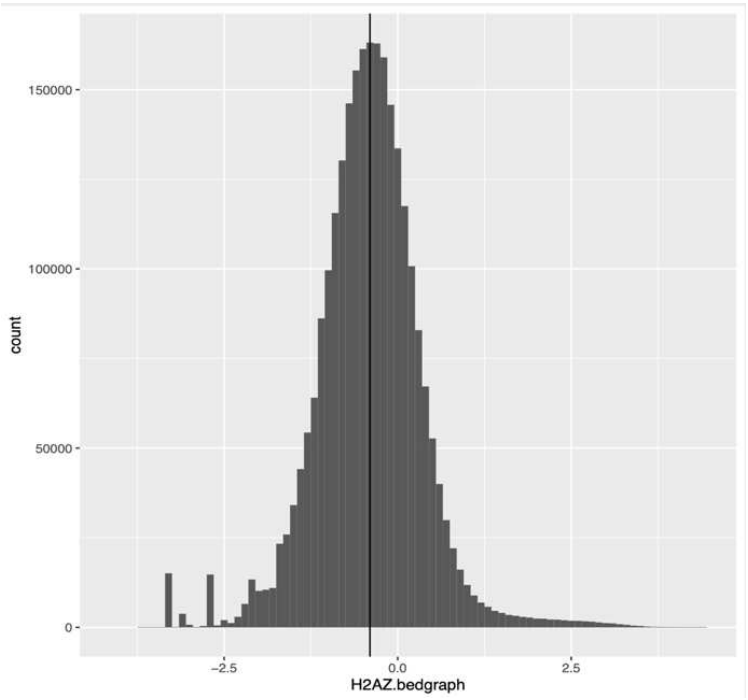


Figure 5. 10. The average signal strength of a histone ChIP-seq data per 1kb bin distribution before normalization. In this plot the bin width is 0.1, the peak of distribution marked by the black vertical line is 0.35 away from 0. So, all average signals will add 0.35 to adjust the peak of it to 0. All downloaded histone ChIP-seq data were repeated the same operation to perform the normalization.

5.3.1.3 The various histone marks enriched around IZs

In Fig. 5.11, we can easily observe the enrichment of DNase I and H3K4me1 at the center of initial zones even in the 3rd and 4th quantiles of IZs with late replication timing (S50) values. There is also an enriched peak signal of H2A.Z at the centers of IZs as expected. As for the H3K79me2 and RNA Pol II-phosphoS2, which are highly correlated with gene transcription, they show double peaks at two sides of initial zone centers. This represents a staggered distribution of transcription and replication which might avoid the conflict between these two biological processes.

Based on the results shown in Figure 5.11 further supports the reliability of the initial zones that we identified. All modification marks show corresponding enrichment around initiation zones as we expected. The enrichment is stronger for ORM IZs in the 1st and 2nd quantile groups. Signal enrichment can still be observed for the IZs in the 3rd quantile group. Analysis with additional histone modification marks enrichment plots can be found in supplementary data.

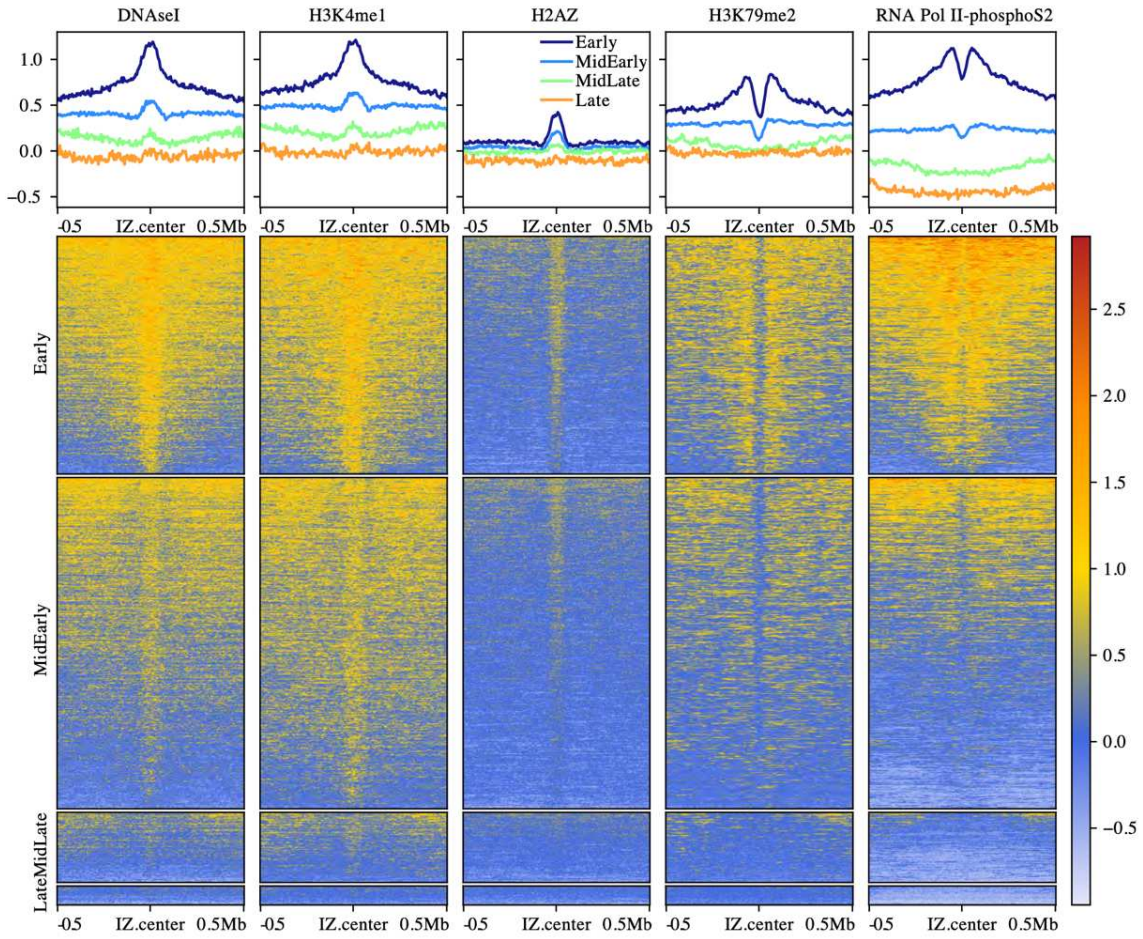


Figure 5. 11. The distribution of DNase I hyper-sensitive sites and histone modification marks around ORM initial zones. On top**,** in the density curves, the four colors (dark blue, light blue, green and orange) represent 4 quantiles as indicated, from the 1st to the 4th represent the group of ORM IZ with replication timing value (S50) from early to late. The heatmaps below show the signal strength distribution for each IZ. Each row in the heatmap represents the signal distribution around one initial zone. The signal strength from strong to weak is based on the color bar from orange to blue as indicated on the right side.

5.3.1.4 The GC enriched at ORM initial zones but not caused by G4 motifs

We calculated the GC content by percentage in each 1 kb bin along the genome and used similar methods to draw the content distribution around initial zones of 4 replication-timing quantile groups (Fig. 5.12). There is also an obvious enrichment of GC-rich sequences at the centers of initial zones. As shown in the heatmap in Figure 5.7, there is also a positive correlation between G4 and ORM track density. Therefore, we tested whether the IZs we identified were indeed enriched with G4 motifs or it's due to their high GC content, by using a GC-content-adjusted background model to avoid detecting G4 enrichment as a trivial consequence of the GC-rich nature of enhancers (where IZs enrich). Of the 737,735 G4 sequences computationally identified in the human genome (Puig Lombardi et al., 2019), 68,585 are found in IZs, which is not more than would be expected by chance, given their GC-rich nature (partial correlation r = -0.0513). Moreover, GC-rich regions (with similar GC% as IZs) outside of IZ are no less likely to contain G4 motifs (131,114 with G4 and 94,596 without) than GC-rich regions within IZs (10,889 with G4 and 7,736 without; chi-square test, p = 0.3227). Therefore, although we find that IZs are GC rich (Fig. 5.12), as previously reported (Xu et al., 2012; Cayrou et al., 2015), they appear to contain G4 motifs as a consequence of their GC richness but do not appear to be enriched for G4 motif DNA sequences.

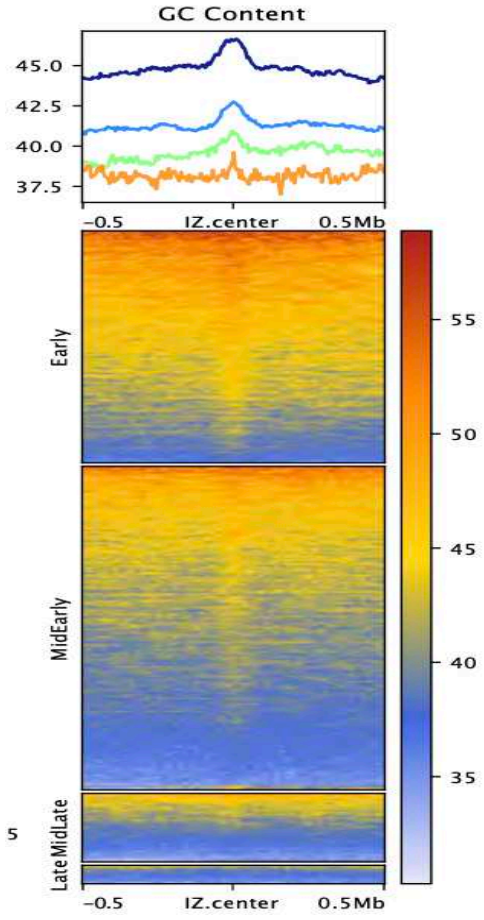


Fig 5.12. The distribution of GC Content around ORM IZs. Similar to Figure 5.10, the GC content enriched in 4 quantiles from early to late.

5.3.1.5 The functional annotation of ORM initial zones

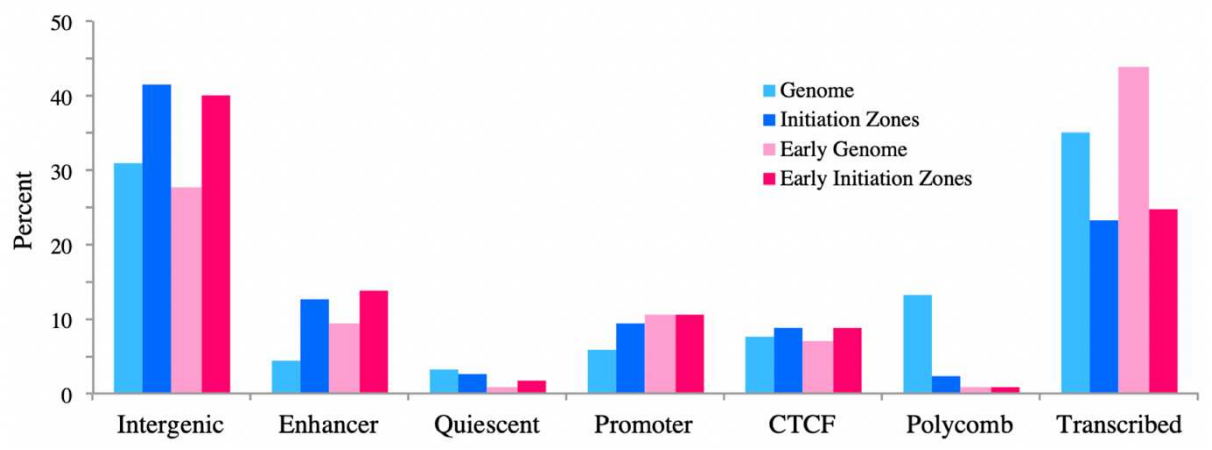


Figure 5. 13. The percentage in various genetic functional regions of ORM initiation zones (Initiation zones and Early Initiation Zones) and control groups (Genome and Early Replicating Genome). The light blue and dark blue represent the genome control and initiation zones, respectively. The pink and red represent early genome control and early initiation zones respectively. Early regions are regions with replication timing value (S50) < = 0.25, and early initiation zones are the ORM initiation zones, the centers of which have a S50 <= 0.25.

For further checking the functional characterization of ORM initial zones, we analyzed the functional annotation of ORM IZ regions and calculated the percentage belong to each type of functional region including Intergenic, Enhancer, Quiescent, Promoter, CTCF, Polycomb, and Transcribed. After comparing our ORM IZs with various chromatin states defined from the ENCODE project (Bernstein et al., 2012; Ernst and Kellis, 2012), we found that they are enriched in enhancers and low activity intergenic regions, and depleted in transcription units and polycomb-repressed chromatin (Fig. 5.13). These observations agree with the enrichment of corresponding histone marks observed in the previous section (Fig. 5.11).

CHAPTER 6

Conclusion and perspectives

We have developed a novel optical method to do replication mapping and origin detection, call Optical Replication Mapping (ORM). ORM technique is the first attempt in the genome-wide mapping of individual replication origins of human cells by a high-throughput, single-molecule, relatedly cheap method. ORM combines a series of advantages. In addition, to map individual replication origins, it can also use to determine the origin fire efficiency to simulate the DNA replication process through normalized signal density. By simulated such a process, we provide strong evidence for supporting the stochastic model of DNA replication timing regulation. The estimation obtained by ORM is out-performing comparing to other origin mapping methods based on population-based data.

At the same time, the experimental cost of this method is lower than nanopore sequencing technology, and the operation process is simple, which mainly depends on the *in vivo* labeling of ongoing replicating regions and optical replication mapping by the Bionano Saphyr platform *in vitro*. The obtained DNA fiber is very solid in terms of sequence length and coverage. There is no need for later assembly analysis, and the ultra-long sequence can also map the DNA fiber to the region containing multiple repeats, which greatly improves the accuracy of mapping. All of this cannot be achieved by the next generation sequencing technology. For a given single flow cell, the average depth can up to 200-300 x coverage of the human genome, and the length of DNA fibers can up to 150 kb ~ 2000 kb (~300 kb on average). In merged 0 min data of all our experiments (4 biological replicates introduced in table 2.1, dataset A.0 contain 4 technical replicates, dataset C.0 and D.0 contain 3 technical replicates), the coverage and average fiber length are 2,550-fold of the human genome and 284 kb, respectively. Recently, a similar technology named high-throughput optical mapping of replicating DNA (HOMARD) using Bionano system combing with *in vitro* replication labeling has also been reported in *Xenopus* egg extracts (De Carli et al., 2018) to study genome-wide analysis of DNA replication.

6.1 Main conclusion

6.1.1 ORM – a future trend in initiation detection: single-molecule, cheap and high-throughput

A major question in the field of DNA replication is to detect replication initiation locations and their fire efficiency. Concerned with their low firing frequency and heterogeneous nature in the selection of replication origin sites between different cells (even within a cell population of a given cell type), which has been frequently reported, the approaches for bulk data analysis have poor sensitivity to detect low fire efficiency initiation events mixed with background noise. So, single-molecule methods are the only way that allows detecting replication initiation events precisely.

Until now, the only two methodologies with characteristics of single molecules are Nanopore sequencing and DNA combing. In terms of experimental cost, it is difficult to perform genome-wide origin detection along the human genome by Nanopore sequencing due to the high cost. It should be noted that the detection of thymine analogs (such as BrdU) incorporated around replication origins needs subsequent complicated data analysis, which has high requirements for algorithms like machine learning (Conrad A. et al., 2019). Some researchers have also reported that no matter the number of datasets (3.8 Gb in yeast, almost 1-fold coverage for human cells) or DNA fiber length (average 32 kb) is so limited. (Georgieva et al., 2019; Müller et al., 2019; Hennion et al., 2020). As for DNA combing, to date, its throughput is only a few hundreds of fibers. So, ORM is the first method that combines single molecule, economical, and high-throughput triple characteristics with super coverage (average 200-fold of the human genome) and ultra-long DNA (average 300 kb).

6.1.2 Direct fire efficiency detection reveals that initiations are not clustered

Through the novel ORM method we developed, we detected the replication initiations and replication process from very early of the S phase (first 2%) to the 90 min after entry of the S phase in aphidicolin-synchronized HeLa cells. At the first 2% of the S phase, we found 977,746 replication tracks distribute across 4,930 broad initiation zones mostly range between 20-50 kb (average length 38 kb). Most initiation events are enriched in the enhancer regions and open chromatin regions associated with DNase I HS and histone modifications like H2AZ and H3K4me1 (Fig 5.10) as reported in previous studies (Cayrou et al., 2015; Ganier et al., 2019; Long et al., 2020; Petryk et al., 2016; Pourkarimi et al., 2016).

More importantly, the advantage of ORM over previous studies is that it can directly calculate the fire efficiency of initiation events since it records all the fibers containing replication signals or not. By measuring both labeling and unlabeled DNA fibers, our ORM data can sensitively detect initiation events with even 0.1% frequency considering a >1000x genome-wide coverage, which makes more initiation events are detected without omission. When more initiation events are fully detected, we are more likely to detect more than one initiation event on the same DNA fiber. This is important because they represent neighbor initiation events from the same cell. Besides that, the super length (average 300 kb) of DNA fibers can increase the probability to collect the fiber with more than one ORM track. Both high sensitivity and ultra-fiber length allow us to accurately measure the correlations between neighbor (or close) initiation events along the human genome. Since some the literature describes the domino-like model in replication timing regulation (Guilbaud et al., 2011; Löb et al., 2016; Sporbert et al., 2002), which suggests the initiation event in one position will probably trigger the initiation in neighbored origin positions along the genome. And the initiation events could be clustered at a distance of 75 kb~150 kb (Blow et al., 2001; Cayrou et al., 2011; Huberman and Tsai, 1973; Jackson and Pombo, 1998; Lebofsky et al., 2006; Marheineke and Hyrien, 2004). Based on probability and statics experience, anything that has a certain probability and happens randomly follows a normal distribution. In the above assumption, the inter-distance between neighbored regions should also follow an expected value in the range 75~150 kb, which fits the definition of normal distribution. Thus, if this domino-like model is correct, the distribution of inter-distance between 2 adjacent initiation sites should be bell-shaped, with a peak at the most frequent inter-event distance 75~150 kb. However, by checking the inter-distance of 2 adjacent ORM track centers in the same DNA fiber (Figure S3.A), we found the

inter-distance distribution is exponential (r = 0.99), inconsistent with any significant clustering of initiation (Birnbaum, 1954). Furthermore, not only the inter-distance distribution didn't show the clustering of initiation events' positions, but the correlation (either positive or negative correlation) between fire efficiency of neighbored IZs sharing the same DNA fiber is also not obvious (Figure S3.B). The result is consistent with the conclusion that human replication initiation sites are not clustered in the early S phase, at least at the very first 2% of S phase.

6.1.3 ORM data support a stochastic model in replication timing regulation

Based on the replication imitation events identified by ORM, there is no sign showing the existence of specific replication sites, which have been reported in some previous studies (Anglana et al., 2003; Besnard et al., 2012; Demczuk et al., 2012; Dijkwel et al., 2002; Hamlin et al., 2008; Tao et al., 2000). Limited by the low labeling efficiency, the current resolution of ORM is ~15 kb. Despite this, there are still 85% IZs containing at least 2 ORM tracks whose centers away from 15 kb (Figure 5.5). This means that they come from different initiation sites within the same IZ. A similar situation also occurs in the Top1 locus, which has been reported as early-firing origins in the human genome (Keller et al., 2002; Tao et al., 2000). There are no isolated, efficient, well-defined replication origins (Fig S5.A) in the Top1 loci neither. It could be due to the low fire efficiency nature of most initiation sites and the poor signal-to-noise ratio in previous studies, which result in the incomprehensive initiation sites mapping across IZs. Based on our ORM data, only about 6% IZs show up in any given cells at the beginning of the S phase.

In addition, 99% early replication IZs of OK-seq are also detected by ORM. Although the majority of initiation tracks are with an early replication timing value (S50 < 0.25), we still detect about 3% IZ center positions within regions with late replication timing (S50 > 0.75). We deliberately compared the ORM signal distribution of unsynchronized cells, and we didn't find the signal enrichment along the genome (Fig 4.8). Besides that, the ORM fire efficiency is highly correlated with replication timing with a 0.8 spearman score (Fig 5.7). Therefore, the few late replication IZs are not false-positive results caused by asynchronous contamination. At the very beginning of the S phase, the few detected initiation events located in late regions recognized by bulk data, suggesting that initiation events occur stochastically with specific probability related to replication timing. This kind of heterogeneous nature consists of a stochastic model in explaining replication timing regulation in yeast (Rhind, 2006; Rhind et al., 2010). Such a model proposes the average replication timing value of bulk data is caused by the frequency of each potential origin selection in human cells. Each potential initiation site corresponds to a probability to fire during the S phase. The higher the probability values be, the more likely the initiation site will be selected to start replication at the early S phase. Thus, on average, the initiation sites with a high probability to fire will have early replication timing values, and *vice versa*. This is not inconsistent with a very small number of cells in the early S phase, choosing an initiation site within regions with late replication timing to initiate. Because the low probability does not mean that it is impossible. The difference from the domino-like model is that the stochastic model believes that the cell's selection of initial sites is independent, i.e. there is no correlation between adjacent initial sites. However, the domino-like model might can describe the phenomenon that happens later on with the progress of replication. Since with ORM, we only detect the initiation event at the very beginning of the S phase and there is a research report the domino-like model works on heterochromatin regions in the late-replication domain (Löb et al., 2016).

To further confirm whether ORM data is consistent with a rigorously stochastic model of replication kinetics, we introduced an analytical model named "Replicon" (Gindin et al., 2014) to simulate the DNA replication process based on normalized ORM signal density in 1 kb resolution as fire efficiency. The result shows an excellent fit with a 0.85 spearman score to experimental data (Fig 5.5). Due to the fact that the Replicon model sets the initiation events occurring independently based on the given probabilities, the fit result may illustrate that there is no need to involve positive or negative correlation between initiation events and this is enough to affect the distribution of replication origin sites. Combining previous research experience, no matter in budding and fission yeast or human Hela cells (de Moura et al., 2010; Kaykov and Nurse, 2015; Patel et al., 2005; Yang et al., 2010), we proposed the stochastic nature of replication timing regulation is conserved across eukaryotes.

6.1.4 Application of ORM technology and perspectives

The ORM technology can not only use to locate the ongoing replication regions but also tell the dynamic information about replication kinetics. Concerned with the effect of aphidicolin on DNA replication, we introduced unperturbed cells and successfully detected the signal polarity in population-based data of asynchronous cells with labeling signal consuming. The direction of ORM signal intensity becomes weaker suggests the orientation of moving replication forks. Similar means have also been reported with both radio- and fluorescently labeled nucleotides (Hennion et al., 2020; Hubermae and Riqos, 1968; Müller et al., 2019). The fork direction is based on about 500-fold coverage fibers and agrees with the replication-fork-directionality profiles obtained by OK-seq (Fig 4.9 and Fig 4.10). The cell line-specific kinetics were also verified by fork direction calculated by ORM data (Fig 4.11). Now, the limit for ORM is not able to tell the individual replication fork direction due to the sparse labeling. In future work, we plan to mark the ongoing replication region successively by double labeling with novel nucleoside analogs. With the success of such approch, we will be able to know all dynamic information for replication forks clearly and locate the position of origins or terminations in asynchronous cells without any physiological effect caused by synchronization. More importantly, we can detect any initiation or termination event across the entire S phase and further verify the stochastic model or domino-like model in the late S phase. This is bound to become the most powerful tool for studying DNA replication program.

Seeing is believing. ORM technology applications are far more than DNA replication initiation detection. We envision ORM being used to measure replication fork speed, replication fork arrests, and reversal, and sister-chromatid exchange, under both normal growth and replication stress conditions. Any research about the biological processes related to DNA synthesis may get conclusive optical image evidence, such as DNA replication, DNA recombination, DNA repair, and even genome instability.

However, the biggest current limitation of ORM is the inadequate labeling, which makes the sparse signals cannot fully marking the entire replication initiation regions, resulting in final IZs at 15 kb resolution. In current single-molecule methods, the two most common thymidine analogs used are BrdU and EdU. However, both of them are incompatible with ORM technology. BrdU detection is a method of immunodetection of incorporated BrdU using anti-BrdU with a fluorophore. Due

to the large molecular weight of the anti-BrdU, it is difficult to transfer into DNA and be effectively detected, so the BrdU detection requires the use of DNase or HCl or heating to cause DNA to denature, which makes DNA turn into a single strand. But the DNA fibers inside the channel of the Bionano platform need to be double strand DNA molecules. As for EdU, it will keep the original double-strand DNA but the downstream protocol Click-iT needs to add copper ions for the catalytic reaction. The copper ions will cut DNA, which makes it impossible to obtain ultra-long DNA molecules. So, in our experimental setting, the experimental labeling method is using fluorescently conjugated dUTP to mark ongoing replication regions by electroporating, which limits the cell lines we can uses. Recently, there are some other nucleoside analogs have been demonstrated to do DNA replication labeling. Though none of them can be applied to ORM at the current stage, we still predict, with the development of nucleoside-labeling technology, ORM can be applied to most cell lines, tissues, and organisms at 1 kb resolution to increase the versatility of ORM technology.

In addition, as for the specific biological question where ORM method or ORM data can use, it can apply to the field of replication initiation control, timing regulation, and any gene function related to the DNA replication process. I will list several possible application directions below.

6.1.4.1 Is initiation really controlled by DNA base composition signature?

Although the current academic community generally believes that there are no definite initiation sites in mammalian cells, there are still some studies pointing out that the existence of a conserved DNA base composition signature can define the replication initiation zone (Ganier et al., 2019). However, the initiation event is detected by SNS-seq in this research. In section 1.3.1, we used to talk about the small nascent DNA that may accumulate at the fork stalling positions. So, SNS-seq may introduce some biological bias for the G4 enrichment in their origin sites. Similarly, one of the main results of the DNA base composition signature in this article is the G-rich element. Thus, the ORM method can be applied to test the authenticity of the conclusion obtained by SNS-seq. In the following, for the convenience of description, we refer to the origin sites containing the conserved DNA base composition signature as “signature-sites”.

Firstly, labeling of ORM has achieved encouraging progress in our teams (Chun-long Chen, UMR3244, Curie Institute, Paris). So, in the future, we might can apply the ORM method to unsynchronized cells to see if the signature-sites are real initiation sites directly. Meanwhile, the other usage of ORM is to detect the positions of stalling forks efficiently by checking the similar replication fork stop position in several DNA fibers. This has higher requirements for DNA fiber’s coverage. Compared with the only two other single-molecule methods: Nanopore and DNA combing, it seems that only ORM coverage can be used to efficiently detect large amount of stalling forks and their corresponding locations. We can check whether the positions with DNA base composition signature and G4 enrichment situation in fork stalling positions are consistent with the reported characteristics in origin sites detected by SNS-seq, even include the SNS-seq origin sites.

We can also remove or copy the region with clustered signature-sites to other or several intergenic regions to test whether the relocation of regions with such signature can trigger the new initiation zone distribution and compare the average ORM fire efficiency between relocated regions and

original regions. The choice of ORM will not bring the biological bias like SNS-seq. At the same time, the single-molecule method can detect the low fire efficiency initiation site to provide more position to verify the correctness of signature-sites.

6.1.4.2 Is initiation cohesin-dependent or not?

A lot of literature has reported that replication initiation often appears in TAD (Topologically associating domains) borders, this is because the high-density arrays of co-occupied CTCF+cohesin binding sites within TAD regions may make replication origin avoid anchoring inside the TAD regions. However, whether the initiation event absolutely depends on the cohesin of the genome region is still under debate. One study reported that the cohesin-mediated genome architecture doesn't define DNA replication Timing domains (Oldach and Nieduszynski, 2019). The latest research proposed the opposite view that cohesin-mediated loop anchors confine the location of human replication origins (Emerson et al., 2021), and maybe cohesin knockdown treatment by small interfering RNA in the previous study can't totally deplete the cohesin effect. Thus, such controversial topics can be fully verified by ORM.

For testing the effect of cohesin on initiation distribution and replication timing, we can compare the ORM signal density of TAD regions between engineered fully Rad21 cohesin degraded condition and wild type condition. If there is obvious ORM signal density difference in TAD regions after knock-down of cohesion mediated loops in G1. We can know the correlation between cohesin and initiation.

The advantage of ORM method in this project is its high sensitivity to detect the low fire efficiency initiation. Because there is no guarantee by any means that the influence of cohesion in the TAD area can be completely eliminated, so the initiation fire efficiency of the TAD area may not be very high, which requires a high DNA fiber coverage and single-molecule strategy, but even only 1%. All of the invitations that happen here can be detected by the ORM method, which cannot be done by other methods.

6.1.4.3 Domino-like model testing in late replicating domain

Reports on the domino-like model are mainly focused on the late replicated region of heterochromatin (Löb et al., 2016). In the end, whether the timing regulation mechanism really exists can be tested with improved ORM double labeling.

First of all, we can detect the replication origins in the late replicating domain by double labeling. Then we can calculate whether the inter-distance of adjacent ORM origins on the same DNA fiber follows the normal distribution with the expected inter-distance between 75~150 kb, the detailed reason in section 6.1.2. Is there a positive correlation in fire efficiency between two adjacent ORM origins on the same DNA fiber? Check the possibility of domino-like existence through mathematical-statistical analysis. If the statistical results show that the domain-like model really exists, combined with the stochastic model in the early S phase in this article, it means that there may be different replication mechanisms in the early and late S phases.

Based on the reported 3D domino effect (Löb et al., 2016), we can speculate that this domino-like timing regulation mechanism may be triggered by some kind of physical interaction of unknown protein related to the domino-like model. At least, one research has pointed out the possible molecule difference between the replication during the early and late S phases. The ChIP-seq of RIF1 clustered only in the late replication region (Foti et al., 2016).

Some other researchers also pointed out that the dormant origins replicated passively by active replication forks emerging from adjacent origins (Burkhart et al., 1995; Ibarra et al., 2008). However, the ATR and CHK1 can block the process of such dormant origins' replication firing by active replication forks, which could be the unknown mechanism inhibits additional origin firing to suppress the potential domino-like model (Moiseeva et al., 2019).

No matter, the Rif1 or some unknown genes that work in the late S phase trigger the domino-like model. Or the domino-like model is an inherent suppressed nature across the entire S phase, but only in the late S phase, the inhibition from ATK and CHK1 will be canceled to make the domino-like model more obvious. In summary, all the exploration for the relationship between the specific gene function and domino-like model can be tested by ORM. Because, firstly, domino-like model point to the active replication initiation event, which can't be through any approach like ChIP-seq, which may trigger some dormant origins mixed with the final result. And the data for verification of the domino-like model is the adjacent initiation sites from the same cells/DNA fibers. The single-molecule characteristics allow us to detect all possible initiation sites comprehensively. This virtually increases the probability of observing multiple initiation sites on the same DNA fiber. Besides that, the advantage of ORM is the only method that has both high coverage and ultra-long DNA fiber, which can provide a large number of initiation sites from the same DNA fibers. Large-scale samples are essential for probability statistics.

Reference

Abdurashidova, G., Radulescu, S., Sandoval, O., Zahariev, S., Danailov, M.B., Demidovich, A., Santamaria, L., Biamonti, G., Riva, S., Falaschi, A., 2007. Functional interactions of DNA topoisomerases with a human replication origin. EMBO J. 26, 998–1009. https://doi.org/10.1038/sj.emboj.7601578

Allemand, J.F., Bensimon, D., Jullien, L., Bensimon, A., Croquette, V., 1997. slides coated with 3-aminopropyltriethoxysilane (Sigma) as described in (3) / stretched on silanized microscope slides by molecular combing (Fig. 2B) (5–8) -8. Biophys. J. 73, 2064–2070.

Anglana, M., Apiou, F., Bensimon, A., Debatisse, M., 2003. Dynamics of DNA replication in mammalian somatic cells: Nucleotide pool modulates origin choice and interorigin spacing. Cell 114, 385–394. https://doi.org/10.1016/S0092-8674(03)00569-5

Beall, E.L., Manak, J.R., Zhou, S., Bell, M., Lipsick, J.S., Botchan, M.R., 2002. Role for a Drosophila Myb-containing protein complex in site-specific DNA replication. Nature 420, 833–837. https://doi.org/10.1038/nature01228

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J.M., Lemaitre, J.M., 2012. Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat. Struct. Mol. Biol. 19, 837–844. https://doi.org/10.1038/nsmb.2339

Birnbaum, A. (Columbia U., 1954. Statistical Methods for Poisson Processes and Exponential Populations Author (s): Allan Birnbaum Source : Journal of the American Statistical Association , Vol . 49 , No . 266 (Jun ., 1954), pp . 254-. J. Am. Stat. Assoc. 49, 254–266.

Blow, J.J., Gillespie, P.J., Francis, D., Jackson, D.A., 2001. Replication origins in Xenopus egg extract are 5-15 kilobases apart and are activated in clusters that fire at different times. J. Cell Biol. 152, 15–25. https://doi.org/10.1083/jcb.152.1.15

Brison, O., Gnan, S., Azar, D., Schmidt, M., Koundrioukoff, S., El-Hilali, S., Jaszczyszyn, Y., Lachages, A.M., Thermes, C., Chen, C.L., Debatisse, M., 2020. Unscheduled origin building in S-phase upon tight CDK1 inhibition suppresses CFS instability. bioRxiv. https://doi.org/10.1101/2020.11.19.390054

Burkhart, R., Schulte, D., Hu, B., Musahl, C., Göhring, F., Knippers, R., 1995. Interactions of Human Nuclear Proteins P1Mcm3 and P1Cdc46. Eur. J. Biochem. 228, 431–438. https://doi.org/10.1111/j.1432-1033.1995.tb20281.x

Cayrou, C., Ballester, B., Peiffer, I., Fenouil, R., Coulombe, P., Andrau, J.C., Van Helden, J., Méchali, M., 2015. The chromatin environment shapes DNA replication origin organization and defines origin classes. Genome Res. 25, 1873–1885. https://doi.org/10.1101/gr.192799.115

Cayrou, C., Coulombe, P., Vigneron, A., Stanojcic, S., Ganier, O., Peiffer, I., Rivals, E., Puy, A., Laurent-Chabalier, S., Desprat, R., Méchali, M., 2011. Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. Genome Res. 21, 1438–1449. https://doi.org/10.1101/gr.121830.111

Chen, C.L., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit, B., D'Aubenton-Carafa, Y., Arneodo, A., Hyrien, O., Thermes, C., 2010. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res. 20, 447–457. https://doi.org/10.1101/gr.098947.109

Chesnokov, I.N., 2007. Multiple Functions of the Origin Recognition Complex. Int. Rev. Cytol. https://doi.org/10.1016/S0074-7696(07)56003-1

Chevereau, G., Audit, B., Zaghloul, L., Aubenton-carafa, Y., Thermes, C., Arneodo, A., 2009. Open chromatin encoded in DNA sequence is the signature of ‘ master ’ replication origins in human cells 37, 6064–6075. https://doi.org/10.1093/nar/gkp631

Chuang, R.Y., Kelly, T.J., 1999. The fission yeast homologue of Orc4p binds to replication origin DNA via multiple AT-hooks. Proc. Natl. Acad. Sci. U. S. A. 96, 2656–2661. https://doi.org/10.1073/pnas.96.6.2656

Comoglio, F., Schlumpf, T., Schmid, V., Rohs, R., Beisel, C., Paro, R., 2015. High-Resolution Profiling of Drosophila Replication Start Sites Reveals a DNA Shape and Chromatin Signature of Metazoan Origins. Cell Rep. 11, 821–834. https://doi.org/10.1016/j.celrep.2015.03.070

De Carli, F., Menezes, N., Berrabah, W., Barbe, V., Genovesio, A., Hyrien, O., 2018. High-Throughput Optical Mapping of Replicating DNA. Small Methods 2, 1800146. https://doi.org/10.1002/smtd.201800146

de Moura, A.P.S., Retkute, R., Hawkins, M., Nieduszynski, C.A., 2010. Mathematical modelling of whole chromosome replication. Nucleic Acids Res. 38, 5623–5633. https://doi.org/10.1093/nar/gkq343

Dellino, G.I., Cittaro, D., Piccioni, R., Luzi, L., Banfi, S., Segalla, S., Cesaroni, M., Mendoza-Maldonado, R., Giacca, M., Pelicci, P.G., 2013. Genome-wide mapping of human DNA-replication origins: Levels of transcription at ORC1 sites regulate origin selection and replication timing. Genome Res. 23, 1–11. https://doi.org/10.1101/gr.142331.112

Demczuk, A., Gauthier, M.G., Veras, I., Kosiyatrakul, S., Schildkraut, C.L., Busslinger, M., Bechhoefer, J., Norio, P., 2012. Regulation of DNA replication within the immunoglobulin heavy-chain locus during B cell commitment. PLoS Biol. 10, 15. https://doi.org/10.1371/journal.pbio.1001360

Diermeier-Daucher, S., Clarke, S.T., Hill, D., Vollmann-Zwerenz, A., Bradford, J.A., Brockhoff, G., 2009. Cell type specific applicability of 5-ethynyl-2′-deoxyuridine (EDU) for dynamic proliferation assessment in flow cytometry. Cytom. Part A 75, 535–546. https://doi.org/10.1002/cyto.a.20712

Dijkwel, P.A., Wang, S., Hamlin, J.L., 2002. Initiation Sites Are Distributed at Frequent Intervals in the Chinese Hamster Dihydrofolate Reductase Origin of Replication but Are Used with Very Different Efficiencies. Mol. Cell. Biol. 22, 3053–3065. https://doi.org/10.1128/mcb.22.9.3053-3065.2002

Dileep, V., Gilbert, D.M., 2018. Single-cell replication profiling to measure stochastic variation in mammalian replication timing. Nat. Commun. 9. https://doi.org/10.1038/s41467-017-02800-w

Eaton, M.L., Galani, K., Kang, S., Bell, S.P., MacAlpine, D.M., 2010. Conserved nucleosome positioning defines replication origins. Genes Dev. 24, 748–753. https://doi.org/10.1101/gad.1913210

Emerson, D., Zhao, P.A., Klein, K., Ge, C., Zhou, L., Sasaki, T., Yang, L., Venvev, S. V., Gibcus, J.H., Dekker, J., Gilbert, D.M., Phillips-Cremins, J.E., 2021. Cohesin-mediated loop anchors confine the location of human replication origins Daniel. bioRxiv 1–25.

Foti, R., Gnan, S., Cornacchia, D., Dileep, V., Bulut-Karslioglu, A., Diehl, S., Buness, A., Klein, F.A., Huber, W., Johnstone, E., Loos, R., Bertone, P., Gilbert, D.M., Manke, T., Jenuwein, T., Buonomo, S.C.B., 2016. Nuclear Architecture Organized by Rif1 Underpins the Replication-Timing Program. Mol. Cell 61, 260–273. https://doi.org/10.1016/j.molcel.2015.12.001

Fragkos, M., Ganier, O., Coulombe, P., Méchali, M., 2015a. DNA replication origin activation in space and time. Nat. Rev. Mol. Cell Biol. https://doi.org/10.1038/nrm4002
Fragkos, M., Ganier, O., Coulombe, P., Méchali, M., 2015b. REVIEWS DNA replication origin activation in space and time. Nat. Publ. Gr. 16, 360–374. https://doi.org/10.1038/nrm4002
Francesco De Carli, 2017. Towards genome-wide , single-molecule analysis of eukaryotic DNA replication To cite this version : HAL Id : tel-01598875 Université Pierre et Marie Curie École doctorale Complexité du Vivant Towards genome-wide , single-molecule analysis of eukaryotic DN.
Ganier, O., Prorok, P., Akerman, I., Méchali, M., 2019. Metazoan DNA replication origins. Curr. Opin. Cell Biol. 58, 134–141. https://doi.org/10.1016/j.ceb.2019.03.003
Gilbert, D.M., 2010. Evaluating genome-scale approaches to eukaryotic DNA replication. Nat. Rev. Genet. https://doi.org/10.1038/nrg2830
Gindin, Y., Meltzer, P.S., Bilke, S., 2014. Replicon : a software to accurately predict DNA replication timing in metazoan cells 5, 1–5. https://doi.org/10.3389/fgene.2014.00378
Gómez, M., Antequera, F., 2008. Overreplication of short DNA regions during S phase in human cells 375–385. https://doi.org/10.1101/gad.445608.clin-dependent
Guilbaud, G., Rappailles, A., Baker, A., Chen, C.L., Arneodo, A., Goldar, A., d'Aubenton-Carafa, Y., Thermes, C., Audit, B., Hyrien, O., 2011. Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. PLoS Comput. Biol. 7. https://doi.org/10.1371/journal.pcbi.1002322
Hamlin, J.L., Mesner, L.D., Lar, O., Torres, R., Chodaparambil, S. V., Wang, L., 2008. A revisionist replicon model for higher eukaryotic genomes. J. Cell. Biochem. 105, 321–329. https://doi.org/10.1002/jcb.21828
Hennion, M., Arbona, J.M., Cruaud, C., Proux, F., Le Tallec, B., Novikova, E., Engelen, S., Lemainque, A., Audit, B., Hyrien, O., 2018. Mapping DNA replication with nanopore sequencing. bioRxiv 1–17. https://doi.org/10.1101/426858
Hennion, M., Arbona, J.M., Lacroix, L., Cruaud, C., Theulot, B., Le Tallec, B., Proux, F., Wu, X., Novikova, E., Engelen, S., Lemainque, A., Audit, B., Hyrien, O., 2020. FORK-seq: replication landscape of the Saccharomyces cerevisiae genome by nanopore sequencing. Genome Biol. https://doi.org/10.1101/2020.04.09.033720
Houchens, C.R., Lu, W., Chuang, R.Y., Frattini, M.G., Fuller, A., Simancek, P., Kelly, T.J., 2008. Multiple mechanisms contribute to Schizosaccharomyces pombe origin recognition complex-DNA interactions. J. Biol. Chem. 283, 30216–30224. https://doi.org/10.1074/jbc.M802649200
Hua, H., Kearsey, S.E., 2011. Monitoring DNA replication in fission yeast by incorporation of 5-ethynyl-2′-deoxyuridine. Nucleic Acids Res. 39. https://doi.org/10.1093/nar/gkr063
Hubermae, J.A., Riqos, A., 1968. On the Mechanism in Mammalian of DNA Replication Chromosomes. J. Mol. Biol. 32, 327–341.
Huberman, J.A., Tsai, A., 1973. Direction of DNA replication in mammalian cells. J. Mol. Biol. 75, 5–8. https://doi.org/10.1016/0022-2836(73)90525-1
Ibarra, A., Schwob, E., Méndez, J., 2008. Excess MCM proteins protect human cells from replicative stress by licensing backup origins of replication. Proc. Natl. Acad. Sci. U. S. A. 105, 8956–8961. https://doi.org/10.1073/pnas.0803978105
Jackson, D.A., Pombo, A., 1998. Replicon clusters are stable units of chromosome structure: Evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. J. Cell Biol. 140, 1285–1295.

https://doi.org/10.1083/jcb.140.6.1285
Karnani, N., Taylor, C.M., Malhotra, A., Dutta, A., 2009. Genomic Study of Replication Initiation in Human Chromosomes Reveals the Influence of Transcription Regulation and Chromatin Structure on Origin Selection. Mol. Biol. Cell 21, 393–404. https://doi.org/10.1091/mbc.E09
Kaykov, A., Nurse, P., 2015. The spatial and temporal organization of origin firing during the S-phase of fission yeast. Genome Res. 25, 391–401. https://doi.org/10.1101/gr.180372.114
Keller, C., Ladenburger, E.M., Kremer, M., Knippers, R., 2002. The origin recognition complex marks a replication origin in the human TOP1 gene promoter. J. Biol. Chem. 277, 31430–31440. https://doi.org/10.1074/jbc.M202165200
Kirstein, N., Buschle, A., Wu, X., Krebs, S., Blum, H., Kremmer, E., Vorberg, I.M., Hammerschmidt, W., Lacroix, L., Hyrien, O., Audit, B., Schepers, A., 2021. Human ORC/MCM density is low in active genes and correlates with replication time but does not delimit initiation zones. Elife 10, 1–30. https://doi.org/10.7554/eLife.62161
Labit, H., Perewoska, I., Germe, T., Hyrien, O., Marheineke, K., 2008. DNA replication timing is deterministic at the level of chromosomal domains but stochastic at the level of replicons in Xenopus egg extracts. Nucleic Acids Res. 36, 5623–5634. https://doi.org/10.1093/nar/gkn533
Langley, A.R., Gräf, S., Smith, J.C., Krude, T., 2016. Genome-wide identification and characterisation of human DNA replication origins by initiation site sequencing (ini-seq). Nucleic Acids Res. 44, 10230–10247. https://doi.org/10.1093/nar/gkw760
Lebofsky, R., Heilig, R., Sonnleitner, M., Weissenbach, J., Bensimon, A., 2006. Santiago M. Di Pietro,* Juan M. Falco ´n-Pe ´rez,* Danie `le Tenza,† Subba R.G. Setty,‡ Michael S. Marks,‡ Grac ¸a Raposo,† and Esteban C. Dell'Angelica*. Mol. Biol. Cell 18, 986–994. https://doi.org/10.1091/mbc.E06
Liu, Y., Lin, Y., Pasero, P., Chen, C., Liu, Y., Lin, Y., Pasero, P., I, C.C.T., Liu, Y., Lin, Y., Pasero, P., Chen, C., 2020. Topoisomerase I prevents transcription-replication conflicts at transcription termination sites To cite this version : HAL Id : hal-03083175 Topoisomerase I prevents transcription-replication conflicts at transcription termination sites. Mol. Cell. Oncol. 00. https://doi.org/10.1080/23723556.2020.1843951
Löb, D., Lengert, N., Chagin, V.O., Reinhart, M., Casas-Delucchi, C.S., Cardoso, M.C., Drossel, B., 2016. 3D replicon distributions arise from stochastic initiation and domino-like DNA replication progression. Nat. Commun. 7. https://doi.org/10.1038/ncomms11207
Lombraña, R., Álvarez, A., Fernández-Justel, J.M., Almeida, R., Poza-Carrión, C., Gomes, F., Calzada, A., Requena, J.M., Gómez, M., 2016. Transcriptionally Driven DNA Replication Program of the Human Parasite Leishmania major. Cell Rep. 16, 1774–1786. https://doi.org/10.1016/j.celrep.2016.07.007
Long, H., Zhang, L., Lv, M., Wen, Z., Zhang, W., Chen, X., Zhang, P., Li, T., Chang, L., Jin, C., Wu, G., Wang, X., Yang, F., Pei, J., Chen, P., Margueron, R., Deng, H., Zhu, M., Li, G., 2020. H2A.Z facilitates licensing and activation of early replication origins. Nature 577, 576–581. https://doi.org/10.1038/s41586-019-1877-9
Macheret, M., Halazonetis, T.D., 2019. Monitoring early S-phase origin firing and replication fork movement by sequencing nascent DNA from synchronized cells. Nat. Protoc. 14, 51–67. https://doi.org/10.1038/s41596-018-0081-y
Macheret, M., Halazonetis, T.D., 2018. Intragenic origins due to short G1 phases underlie oncogene-induced DNA replication stress. Nature 555, 112–116.

https://doi.org/10.1038/nature25507
Marheineke, K., Hyrien, O., 2004. Control of replication origin density and firing time in Xenopus egg extracts. Role of a caffeine-sensitive, ATR-dependent checkpoint. J. Biol. Chem. 279, 28071–28081. https://doi.org/10.1074/jbc.M401574200
Masai, H., Matsumoto, S., You, Z., Yoshizawa-Sugata, N., Oda, M., 2010. Eukaryotic chromosome DNA replication: Where, when, and how? Annu. Rev. Biochem. 79, 89–130. https://doi.org/10.1146/annurev.biochem.052308.103205
Mesner, L.D., Valsakumar, V., Cieslik, M., Pickin, M., Hamlin, J.L., Bekiranov, S., 2013. Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early- and late-firing origins. Genome Res. 23, 1774–1788. https://doi.org/10.1101/gr.155218.113
Mesner, L.D., Valsakumar, V., Karnani, N., Dutta, A., Hamlin, J.L., Bekiranov, S., 2011. Bubble-chip analysis of human origin distributions demonstrates on a genomic scale significant clustering into zones and significant association with transcription (Genome Research (2011) 21, (377-389)). Genome Res. https://doi.org/10.1101/gr.111328.110.review
Moiseeva, T.N., Yin, Y., Calderon, M.J., Qian, C., Schamus-Haynes, S., Sugitani, N., Osmanbeyoglu, H.U., Rothenberg, E., Watkins, S.C., Bakkenist, C.J., 2019. An ATR and CHK1 kinase signaling mechanism that limits origin firing during unperturbed DNA replication. Proc. Natl. Acad. Sci. U. S. A. 116, 13374–13383. https://doi.org/10.1073/pnas.1903418116
Mukherjee, B., Tomimatsu, N., Burma, S., 2015. Immunofluorescence-based methods to monitor DNA end resection. Stress Responses Methods Protoc. 1292, 67–75. https://doi.org/10.1007/978-1-4939-2522-3
Mukhopadhyay, R., Lajugie, J., Fourel, N., Selzer, A., Schizas, M., Bartholdy, B., Mar, J., Lin, C.M., Martin, M.M., Ryan, M., Aladjem, M.I., Bouhassira, E.E., 2014. Allele-Specific Genome-wide Profiling in Human Primary Erythroblasts Reveal Replication Program Organization 10. https://doi.org/10.1371/journal.pgen.1004319
Müller, C.A., Boemo, M.A., Spingardi, P., Kessler, B.M., Kriaucionis, S., Simpson, J.T., Nieduszynski, C.A., 2019. Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. Nat. Methods 16, 429–436. https://doi.org/10.1038/s41592-019-0394-y
Nieduszynski, C.A., 2016. Global and local regulation of replication origin activity, in: The Initiation of DNA Replication in Eukaryotes. pp. 105–122. https://doi.org/10.1007/978-3-319-24696-3_6
Oldach, P., Nieduszynski, C.A., 2019. Cohesin-mediated genome architecture does not define DNA replication timing domains. Genes (Basel). 10. https://doi.org/10.3390/genes10030196
Patel, P.K., Arcangioli, B., Baker, S.P., Bensimon, A., Rhind, N., 2005. DNA Replication Origins Fire Stochastically in Fission Yeast. Mol. Biol. Cell 17, 308–316. https://doi.org/10.1091/mbc.E05
Petryk, N., Kahli, M., D'Aubenton-Carafa, Y., Jaszczyszyn, Y., Shen, Y., Silvain, M., Thermes, C., Chen, C.L., Hyrien, O., 2016. Replication landscape of the human genome. Nat. Commun. 7, 1–13. https://doi.org/10.1038/ncomms10208
Picard, F., Cadoret, J.C., Audit, B., Arneodo, A., Alberti, A., Battail, C., Duret, L., Prioleau, M.N., 2014. The Spatiotemporal Program of DNA Replication Is Associated with Specific

Combinations of Chromatin Marks in Human Cells. PLoS Genet. 10. https://doi.org/10.1371/journal.pgen.1004282
Pope, B.D., Hiratani, I., Gilbert, D.M., 2011. mammalian development 18, 127–136. https://doi.org/10.1007/s10577-009-9100-8.Domain-wide
Pourkarimi, E., Bellush, J.M., Whitehouse, I., 2016. Spatiotemporal coupling and decoupling of gene transcription with DNA replication origins during embryogenesis in C. elegans. Elife 5, 1–12. https://doi.org/10.7554/elife.21728
Remus, D., Beall, E.L., Botchan, M.R., 2004. DNA topology, not DNA sequence, is a critical determinant for Drosophila ORC-DNA binding. EMBO J. 23, 897–907. https://doi.org/10.1038/sj.emboj.7600077
Rhind, N., 2006. C O M M E N TA R Y DNA replication timing : random thoughts about origin firing 8, 1313–1316.
Rhind, N., Yang, S.C.H., Bechhoefer, J., 2010. Reconciling stochastic origin firing with defined replication timing. Chromosom. Res. 18, 35–43. https://doi.org/10.1007/s10577-009-9093-3
Schaarschmidt, D., Baltin, J., Stehle, I.M., Lipps, H.J., Knippers, R., 2004. An episomal mammalian replicon: Sequence-independent binding of the origin recognition complex. EMBO J. 23, 191–201. https://doi.org/10.1038/sj.emboj.7600029
Sequeira-Mendes, J., Díaz-Uriarte, R., Apedaile, A., Huntley, D., Brockdorff, N., Gómez, M., 2009. Transcription initiation activity sets replication origin efficiency in mammalian cells. PLoS Genet. 5. https://doi.org/10.1371/journal.pgen.1000446
Sporbert, A., Gahl, A., Ankerhold, R., Leonhardt, H., Cardoso, M.C., 2002. DNA polymerase clamp shows little turnover at established replication sites but sequential de novo assembly at adjacent origin clusters. Mol. Cell 10, 1355–1365. https://doi.org/10.1016/S1097-2765(02)00729-3
Sugimoto, N., Maehara, K., Yoshida, K., Ohkawa, Y., Fujita, M., 2018. Genome-wide analysis of the spatiotemporal regulation of firing and dormant replication origins in human cells. Nucleic Acids Res. 46, 6683–6696. https://doi.org/10.1093/nar/gky476
Tao, L., Dong, Z., Leffak, M., Zannis-Hadjopoulos, M., Price, G., 2000. Major DNA replication initiation sites in the c-myc locus in human cells. J. Cell. Biochem. 78, 442–457. https://doi.org/10.1002/1097-4644(20000901)78:3<442::AID-JCB9>3.0.CO;2-1
Theis, J.F., Yang, C., Schaefer, C.B., Newlon, C.S., 1999. DNA sequence and functional analysis of homologous ARS elements of Saccharomyces cerevisiae and S. carlsbergensis. Genetics 152, 943–952. https://doi.org/10.1093/genetics/152.3.943
Toya, M., Iino, Y., Yamamoto, M., 1999. Fission yeast Pob1p, which is homologous to budding yeast Boi proteins and exhibits subcellular localization close to actin patches, is essential for cell elongation and separation. Mol. Biol. Cell 10, 2745–2757. https://doi.org/10.1091/mbc.10.8.2745
Tuduri, S., Tourrière, H., Pasero, P., 2010. Defining replication origin efficiency using DNA fiber assays. Chromosom. Res. 18, 91–102. https://doi.org/10.1007/s10577-009-9098-y
Valenzuela, M.S., Chen, Y., Davis, S., Yang, F., Walker, R.L., Bilke, S., Lueders, J., Martin, M.M., Aladjem, M.I., Massion, P.P., Meltzer, P.S., 2011. Preferential localization of human origins of DNA replication at the 5′-ends of expressed genes and at evolutionarily conserved DNA sequences. PLoS One 6. https://doi.org/10.1371/journal.pone.0017308
Vashee, S., Cvetic, C., Lu, W., Simancek, P., Kelly, T.J., Walter, J.C., 2003. Sequence-independent DNA binding and replication initiation by the human origin recognition complex. Genes Dev. 17, 1894–1908. https://doi.org/10.1101/gad.1084203

Vouzas, A.E., Gilbert, D.M., 2021. Mammalian DNA Replication Timing. Cold Spring Harb. Perspect. Biol. a040162. https://doi.org/10.1101/cshperspect.a040162
Wang, W., Klein, K., Proesmans, K., Yang, H., Marchal, C., Zhu, X., Borman, T., Hastie, A., Weng, Z., Bechhoefer, J., Chen, C., Gilbert, D.M., Rhind, N., 2020. Genome-Wide Mapping of Human DNA Replication by Optical Replication Mapping Supports a Stochastic Model of Eukaryotic Replication.
Weber, J.L., Myers, E.W., 1997. Human Whole-Genome Shotgun Sequencing 401–409.
Yang, S.C.H., Rhind, N., Bechhoefer, J., 2010. Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. Mol. Syst. Biol. 6, 1–13. https://doi.org/10.1038/msb.2010.61

Supplementary

The detailed user manual of all developed jar packages for genetic location identification via Bionano

Prerequisites

1. JDK version 1.8
2. R is required for peak calling
3. About 7GB of available RAM
4. Steps in this document assumes that your current working directory is 'example' and that the 'replicon' executable is in the parent directory.

Introduction for jar package

This is an extension software analyzing data based on Bionano high throughput single-molecule imaging platform determination of any biological process associated with fluorescently labeled DNA, for example: DNA replication, DNA recombination, DNA repair. It provides the analysis on each single molecule as well as the frequency occurred in population-based data along the genome and the given genomic positions of interest.

Notice

Please pay attention most of the packages are jar packages and the packages with. R in Package Name column are R packages

Table S1 Package information summary table

Package Name	Function	input	output
AllRawDataRefining	Summarize the 4 kinds of original input files Calculate the labeling signal position	.bnx; .xmp; .qcmap; .r cmap	.txt
GenerateGTF_ByAllDataRefining_Reformat	Generate a gtf-like files for data visualization by IGV	.txt	.gtf
GetFiberCoordinate_ByTXT	Generate a bed file to record all DNA fibers coordinate in .txt file	.txt	.bed
GetRedflagNumberInSlidingWindow_ByAll DataRefining	Generate sliding bin bed files to record the signal density along the genome	.txt	.bed
GetRedflagNumberInAdjacentWindow_ByA llDataRefining	Generate adjacent bin bed files to record the signal density along the genome	.txt	.bed
GetNewSegmentation_AddSoloSignal	Cluster close ORM signal to segment generate the bed file to record the location of segment	.txt	.bed The distance between signals and primary segments
GMM.R	R package to calculate proper cutoff for segmentation by Gaussian Mixed Model	The distance between signals and primary segments	The plot of Gaussian Mixed Model distribution

Add_FDI_ToSegment	Calculate the FDI to reflect the signal polarity and annotated to the bed file	.bed , .txt	.bed with FDI
GetS50Timing	Add the corresponding S50 annotation to the center of region in bed files	.bed without S50 The raw data for S50 along the genome	.bed with S50
Add_DeltaRFD_ToSegment	Add the corresponding DeltaRFD annotation to the center of region in bed files	.bed The raw data for DeltaRFD along the genome	.bed with DeltaRFD
Calculate_FDI_RFD	Calculate the FDI_RFD based on the FDI within the adjacent window along the genome	The chromosome length .bed with FDI	.bed record adjacent bins along the genome with FDI
LOESS.R	Smooth the raw ORM signal density distribution in adjacent or sliding bin along the genome	.bed with raw ORM signal density	.bed with smooth value of ORM signal density
Multi_peakcalling.R	Call IZs based on the smooth ORM signal density in multi replicates	.bed with smooth value of ORM signal density	.bed to record initial zones
GetMappedCount	Count the overlapped region number from one bed file to a target bed file	The target bed file with overlapped regions got by bedtools	The target bed file with overlapped region number
Abstract_HotBin.R	This R package is used to set the proper cutoff value for hot dots and pick out the hot spot regions by plot of signal distribution.	The bed file got by jar package GetNewSegmentation _AddSoloSignal	The plot for the signal number distribution in narrow bins along the genome.
FilterHotDot	Use this script to pick out all green mapping signals' and red labeling signals' intensity to check the signal intensity distribution difference inside and outside the hot spots' regions. And get the .TXT file after hot spots filtering	.txt .bed record hot spot regions	.txt without hot spot .txt with green and red signal intensity inside and outside hot spot regions
SignalCompare.R	Compare the green and red signals distribution within hot spot regions to prove the hot spot is false positive red signal	.txt with green and red signal intensity inside and outside hot spot regions	The plot for the signal intensity comparison inside and outside the hot spot regions

Supplemental Mathematical Methods (from our manuscript available on bioRvix, Wang et al., 2020)

Modeling the Signal-Intensity Distribution

The intensity of a signal is directly proportional to the number, n, of detected photons. Its probability distribution p(n) results from a combination of two processes: the number of photons coming from each fluorophore and the number of fluorophores inside each resolution-limited region measured. If we assume that the incorporation of fluorophores happens independently, both of these processes are Poisson distributed. The number of photons coming from each fluorophore is Poisson distributed with (unknown) parameter λp. Therefore, if there are N fluorophores in the measured region, the number of photons is Poisson distributed with parameter $N\lambda p$. On the other hand, the number of fluorophores N is Poisson distributed with

parameter Λf . Therefore, the distribution of the number of photons is given.

$$p(n) = \sum_{N=0}^{\infty} \frac{\mathrm{e}^{-\Lambda_f} {\Lambda_f}^N}{N!} \frac{\mathrm{e}^{-N\lambda_p} (N\lambda_p)^n}{n!} . \tag{1}$$

One can simplify this expression as one expects the number of photons (and therefore λp) to be large. Therefore, we can use the Stirling approximation [1],

$$n! \approx \exp\left(n\ln n - n + \frac{1}{2}\ln n + c_0 + \frac{1}{12n} + O(n^{-3}) \right), \tag{2}$$

Where c0 is a constant, to rewrite this to

$$p(n) = \sum_{N=0}^{\infty} \frac{\mathrm{e}^{-\Lambda_f - N\lambda_p} {\Lambda_f}^N}{N!} \mathrm{e}^{n\ln\frac{N\lambda_p}{n} + n + c_0 + \frac{1}{12n}} . \tag{3}$$

The signal intensity x is proportional to the number of photons, x = cn, with an unknown proportionality coefficient,

$$p(x) = \frac{p(cn)}{c} . \tag{4}$$

In experimental data, one cannot determine p(x) for small x due to background signals. Therefore, we need to add a renormalisation constant, a, in which we can absorb the prefactor exp(−Λf + c0), to get

$$p(x) = a \sum_{N=0}^{\infty} \frac{{\Lambda_f}^N}{N!} \exp\left(-N\lambda_p + n\ln\frac{N\lambda_p}{cx} + cx + \frac{1}{12cx} \right) . \tag{5}$$

We now have four unknown parameters: a, Λf , λp and c. These were found via a fit using gnuplot's standard fitting procedure (http://www.gnuplot.info), which gives

$$\begin{aligned} &a = 0.0435 \pm 0.0005, \qquad &&\Lambda f = 0.429 \pm 0.004, \\ &\lambda p = 14.47 \pm 0.09, \qquad &&c = 0.0660 \pm 0.0004. \end{aligned} \tag{6}$$

Probability Distribution of Intersignal Distances

We seek the intersignal distance distribution, pA(l). First, note that

$$p_\ell(\ell) \sim \int_0^\infty \mathrm{d}x\, p(x, x+\ell) = \int_0^\infty \mathrm{d}x\, p(x)\, p(x+\ell|x) , \tag{7}$$

where p(x, x + l) is the joint probability to find one signal at position x and another at x + l, without any signal in between them. pA(l) then averages this quantity over all start positions x. We then express the joint probability of two events as the probability of the first times the probability that the second happens, given the first. The result is the probability to find an intersignal distance of l anywhere along the (semi-infinite) genome segment. We also assume an exponentially decreasing amount of label, which implies an exponentially decreasing incorporation rate:

$$r(x) = \frac{R}{c}\,\mathrm{e}^{-\frac{x}{c}}, \tag{8}$$

where c is the genome distance over which the signal-incorporation rate de- creases by a factor $e^{-1} \approx 0.37$ and c/R is the average distance between two signals at t = 0 (i.e., in the absence of depletion). If the fork speed is v, then c/v is the time it takes for labeled nucleotide concentration to decrease by 37%. If we assume that the nucleotide concentration correlates with signal probability, then the probability to see a signal at position x is also given by

$$p(x) = \frac{R}{c}\,\mathrm{e}^{-\frac{x}{c}}\,. \tag{9}$$

Furthermore, one can check that

$$p(x+\ell|x) = p(\text{No Signal between } x \text{ and } x+\ell)\cdot p(x+\ell)\,, \tag{10}$$

With

$$p(\text{No Signal between } x \text{ and } x+\ell) = \mathrm{e}^{-\int_x^{x+\ell}\mathrm{d}x_0\,\frac{r}{c}\cdot\exp\left(-\frac{x_0}{c}\right)} \tag{11}$$

And

$$p(x+\ell) = \frac{R}{c}\,\mathrm{e}^{-\frac{(x+\ell)}{c}}\,. \tag{12}$$

Therefore, one gets

$$p_\ell(\ell) \sim \int_0^\infty \mathrm{d}x\;\frac{R^2}{c^2}\,\mathrm{e}^{-\frac{2x+\ell}{c}-\int_x^{x+\ell}\mathrm{d}x_0\,\frac{r}{c}\cdot\exp\left(-\frac{x_0}{c}\right)}\,. \tag{13}$$

This integral was approximated using Maple (https://www.maplesoft.com) leading to the final result,

$$p_\ell(\ell) \sim \frac{\mathrm{e}^{-R}\left(\exp\left(\mathrm{e}^{-\frac{\ell}{c}}R\right)R + \mathrm{e}^{R+\frac{\ell}{c}} - \exp\left(\mathrm{e}^{-\frac{\ell}{c}}R + \frac{\ell}{c}\right)(1+R)\right)}{c\left(\mathrm{e}^{\frac{\ell}{c}}-1\right)^2}\,. \tag{14}$$

Implicitly, the model above assumes each fiber samples just a single fork whose origin is at x = 0. Then x is the distance a fork has traveled to the right when the labeled nucleotide (signal) is incorporated. What about more complicated scenarios that a fiber might have? For a single fork moving to the left, the result still holds, as the data reports unsigned intersignal distances. Furthermore, numerical results show that the distribution still approximately holds for fibers with multiple forks from neighboring origins.

Inferring the Position of Initiation

In this section, we describe a method to infer the position at which replication has initiated, given an observed pattern of signals. Assume that one has a segment with signals at positions x1, x2, ..., xn (we set x1 < x2 < ... < xn). If we assume that the segment was initiated at position y, then the probability to observe signals at positions x1, x2, ..., xn, is given by

$$p(\{x_1, x_2, ..., x_n\}|y) \sim \mathrm{e}^{-\int_0^\infty \mathrm{d}\tau\, \lambda(\tau)} \prod_{j=1}^{n} \lambda(|x_j - y|)\,, \tag{15}$$

where λ(x) is the probability to label at a distance x from the initiation,

$$\lambda(x) = R_0\, \mathrm{e}^{-\frac{x}{\ell}}, \tag{16}$$

R0 and l being the label and depletion fit parameters. To estimate the position of y, we can now do a maximum-likelihood estimation,

$$\hat{y} = \mathrm{argmax}_y\, p(\{x_1, x_2, ..., x_n\}|y) = \mathrm{argmax}_y\, \mathrm{e}^{-\frac{\sum_i |x_i - y|}{\ell}}\,. \tag{17}$$

If there is an odd number of signals, then this optimization gives

$$\hat{y} = x_{\frac{n+1}{2}}\,, \tag{18}$$

and if there is an even number of signals, the solution is degenerate and can be anything between $\mathrm{X}\frac{n}{2}$ and $\mathrm{X}\frac{n}{2} + 1$ For our calculation, we set

$$\hat{y} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}\,. \tag{19}$$

We estimated the uncertainty on this estimator by doing 105 simulations using custom C code (available by request). With the correct fit parameters, this gives us a standard deviation of

$$\sigma_{\hat{y}} = 14.2 \pm 0.1\,\mathrm{kb}\,. \tag{20}$$

Estimating the Initiation Event Labeling Efficiency

Using the analysis in Sections 1, 2 and 3, we can estimate the frequency with which an early initiation event will incorporate at least one label and thus be identified by ORM. We begin by

noting that the incorporation rate of at position x of a replication that started at t = 0 is equal to

$$r(x) = r_0 \, \mathrm{e}^{-\frac{x}{c}} \, . \tag{21}$$

This means that at time t, this rate is

$$r(t) = r_0 \, \mathrm{e}^{-\frac{vt}{c}} \, . \tag{22}$$

As r(t) is independent of when the initiation started, one can see that r(x) for an initiation that started at time t0 is given by

$$r(x) = r_0 \, \mathrm{e}^{-\frac{t_0 v + x}{c}} \, . \tag{23}$$

The probability to not get any signals within a distance x0 from the initiation is then

$$\exp\left(-\int_0^{x_0} \mathrm{d}x \, r\,(x)\right) = \exp\left(-r_0 c \left(\mathrm{e}^{-\frac{t_0 v}{c}} - \mathrm{e}^{-\frac{t_0 v + x_0}{c}}\right)\right). \tag{24}$$

Setting v=1.65 kb/min (replication fork rate, from Figure 1C), x0=15 kb (the nominal resolution of ORM from Eq. 20), r0=1/3.8 kb (the initial labeling rate, from Figure S2A) and c=99 kb (Figure S2b; note that the 75 kb reported there is c in base 2, whereas 99 kb used here is c in base e) and assuming that the initiations happen uniformly in early S phase, one estimates that the probability to see zero signals within the first 15 kb of an initiation is 9.5%.

Distribution of Signals within Initiation Zones

Consider an initiation zone of length L. We are interested in determining the distribution of initiations inside the initiation zone. Here, we will consider two extreme cases. The first possibility is that the initiation always happens at a single point at the center of the IZ. The second possibility is that the initiation happens with equal probability everywhere along the IZ.

If the initiation always happens at the center of the IZ, then the probability that a signal is incorporated at the center of the IZ, pc, and the probability that a signal is incorporated at the end of an IZ, pe, are related via

$$\frac{p_e}{p_c} = \exp\left(-\frac{L}{2l}\right), \tag{25}$$

where l is the depletion length. On the other hand, if the initiation happens everywhere with equal probability, then the probability to have a signal at an end of the IZ is given by

$$p_e = \frac{R}{l}\int_0^L \mathrm{d}x \, \mathrm{e}^{-\frac{x}{l}} = R\left(1 - \mathrm{e}^{-\frac{L}{l}}\right), \tag{26}$$

while the probability to have a signal at the center of the IZ is given by

$$p_c = \frac{R}{l}\int_0^L \mathrm{d}x\, \mathrm{e}^{-\frac{\left|x-\frac{L}{2}\right|}{l}} = 2R\left(1-\mathrm{e}^{-\frac{L}{2l}}\right), \qquad (27)$$

which leads to

$$\frac{p_e}{p_c} = \frac{1-\mathrm{e}^{-\frac{L}{l}}}{2(1-\mathrm{e}^{-\frac{L}{2l}})}. \qquad (28)$$

To test whether one of these two models fits the data, we calculate the number of signals within 5 kb of the left end of an IZ (Ne) and the number of signals within 5 kb of the center of the IZ. One then expects

$$\frac{N_e}{N_c} = \frac{p_e}{p_c}. \qquad (29)$$

Therefore, we compare Ne/Nc with Eqs. (25) and (27), where we have deter- mined l from the intersignal distance, $l = 10^5$.

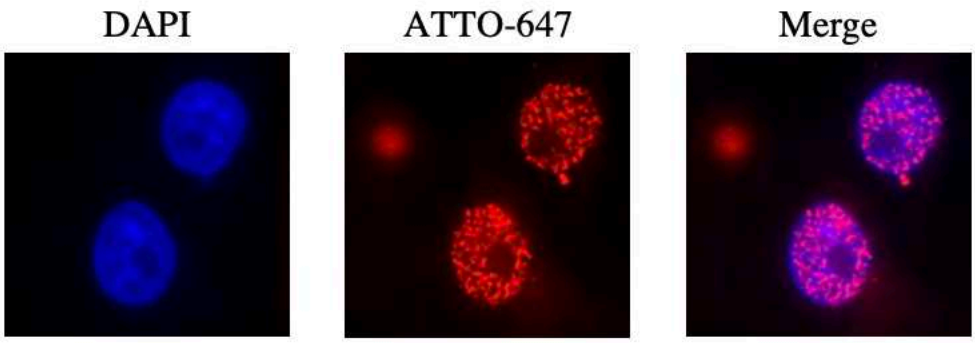


Fig S 1.A Micrographs of representative HeLa cells electroporated with ATTO-647-dUTP during an aphidicolin arrest, released, allowed to recover overnight, and fixed. The left panel is stained with DAPI, the middle panel visualizes the incorporated fluorescent nucleotide, and the right panel is a merger of the two channels.

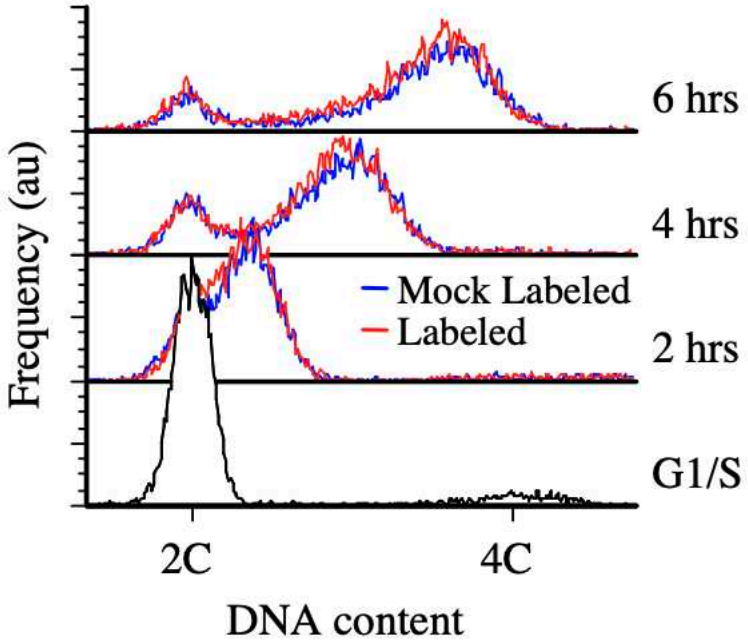


Fig S 1.B Flow cytometry analysis of S-phase progression after ATTO-647-dUTP electroporation.

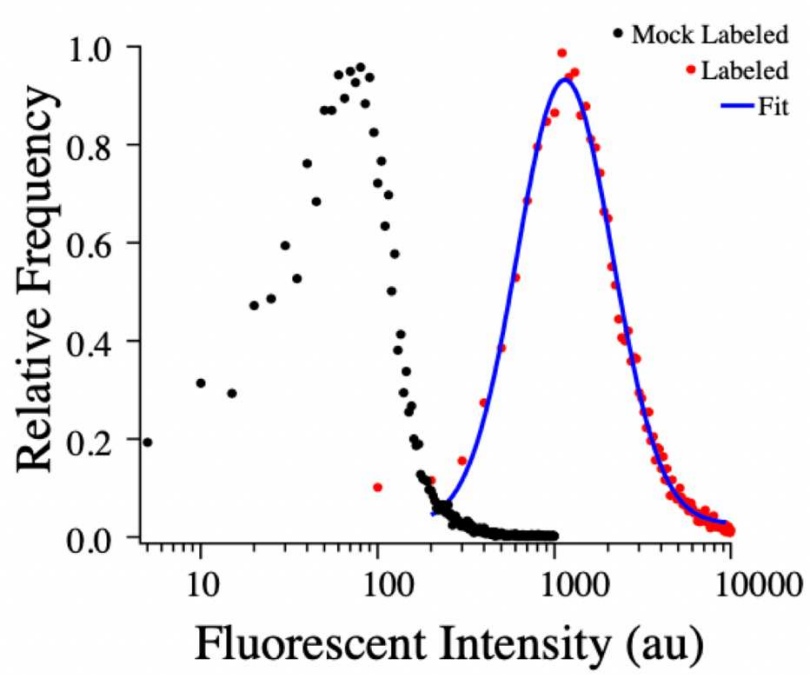


Fig S 1.C Flow cytometry analysis of ATTO-647-dUTP uptake. Cells arrested in aphidicolin at the beginning of S phase were electroporated with ATTO-647-dUTP, or mock electroporated, incubated on ice and analyzed by flow cytometry. The distribution of labeled cells was fit with a log-normal distribution with a mean of 1141 ± 7.7 and a coefficient of variation of 0.88 ± 0.01.

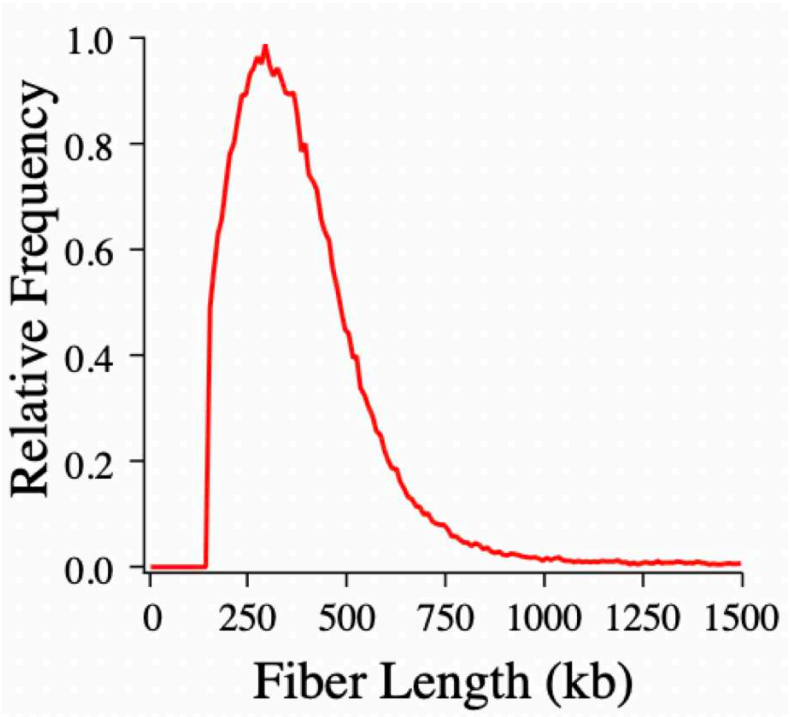


Fig S 1.D Flow cytometry analysis of ATTO-647-dUTP uptake. Cells arrested in aphidicolin at the beginning of S phase were electroporated with ATTO-647-dUTP, or mock electroporated, incubated on ice and analyzed by flow cytometry. The distribution of labeled cells was fit with a log-normal distribution with a mean of 1141 ± 7.7and a coefficient of variation of 0.88 ± 0.01.

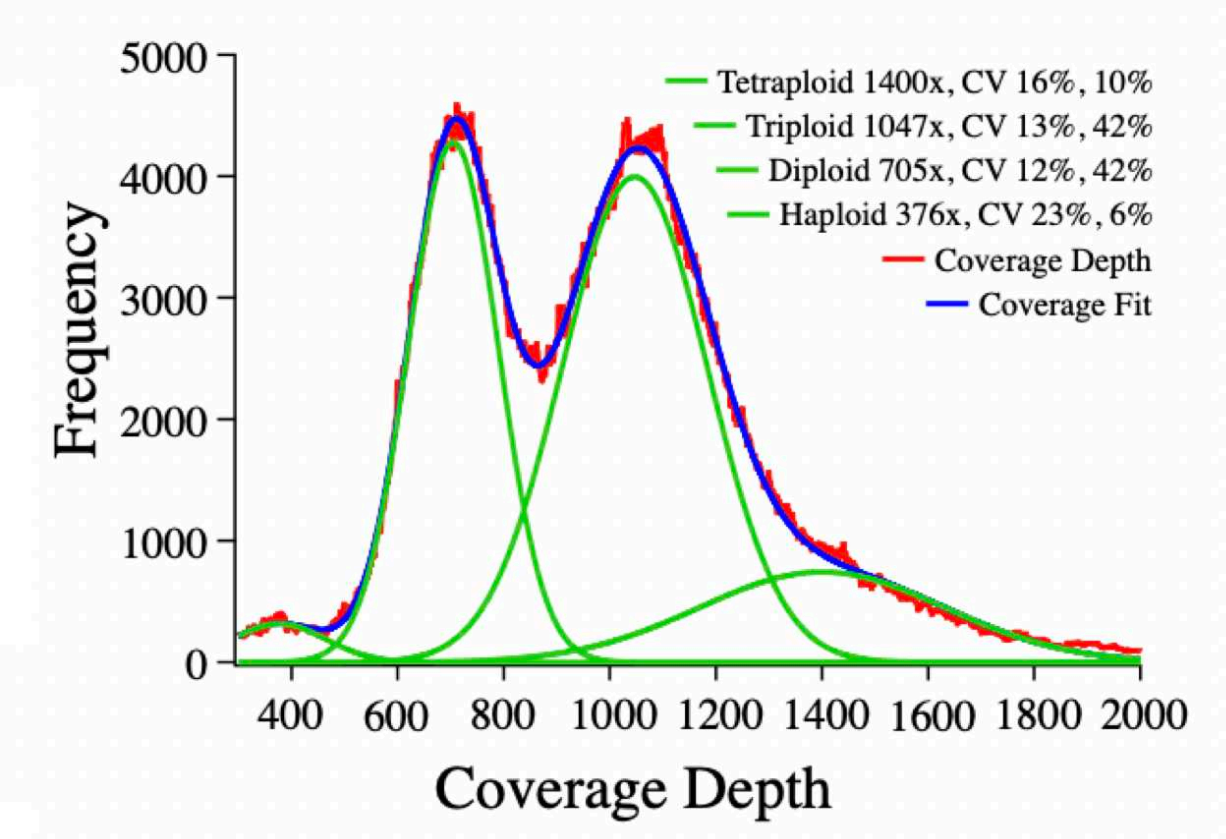


Fig S1.E. The depth of genome coverage in the combined 0-minute dataset in 1 kb bins. The aneuploid character of the HeLa genome is evident in the distribution of coverage into four peaks corresponding to the haploid, diploid, triploid and tetraploid regions of the genome. The coverage data was fit with four Gaussian curves with coverage maxima at about 376x (haploid), 705x (diploid), 1047x (triploid) and 1400x (tetraploid, which was fixed at 1400x, because the unconstrained fit had a very large variation). The individual Gaussians and the complete fit are shown. The coefficients of variation of the individual Gaussians and the percent of the genome inferred to have that ploidy are shown in the legend.

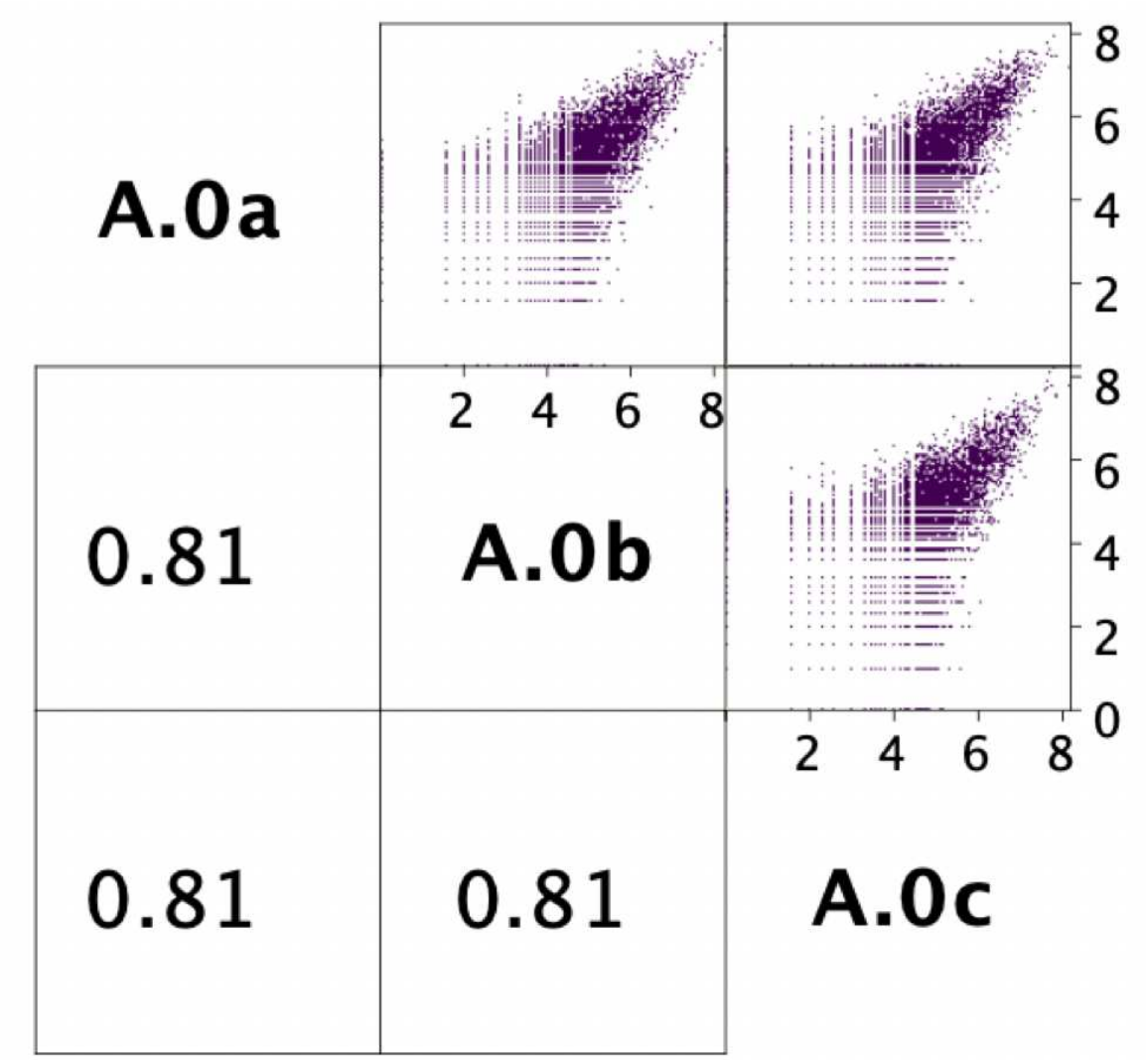


Fig S1.F The correlation of labeling between the three biological replicates that make up the C.0 0-minute dataset. The number of signals in each 10 kb bin across the genome is plotted and the correlation coefficient is reported.

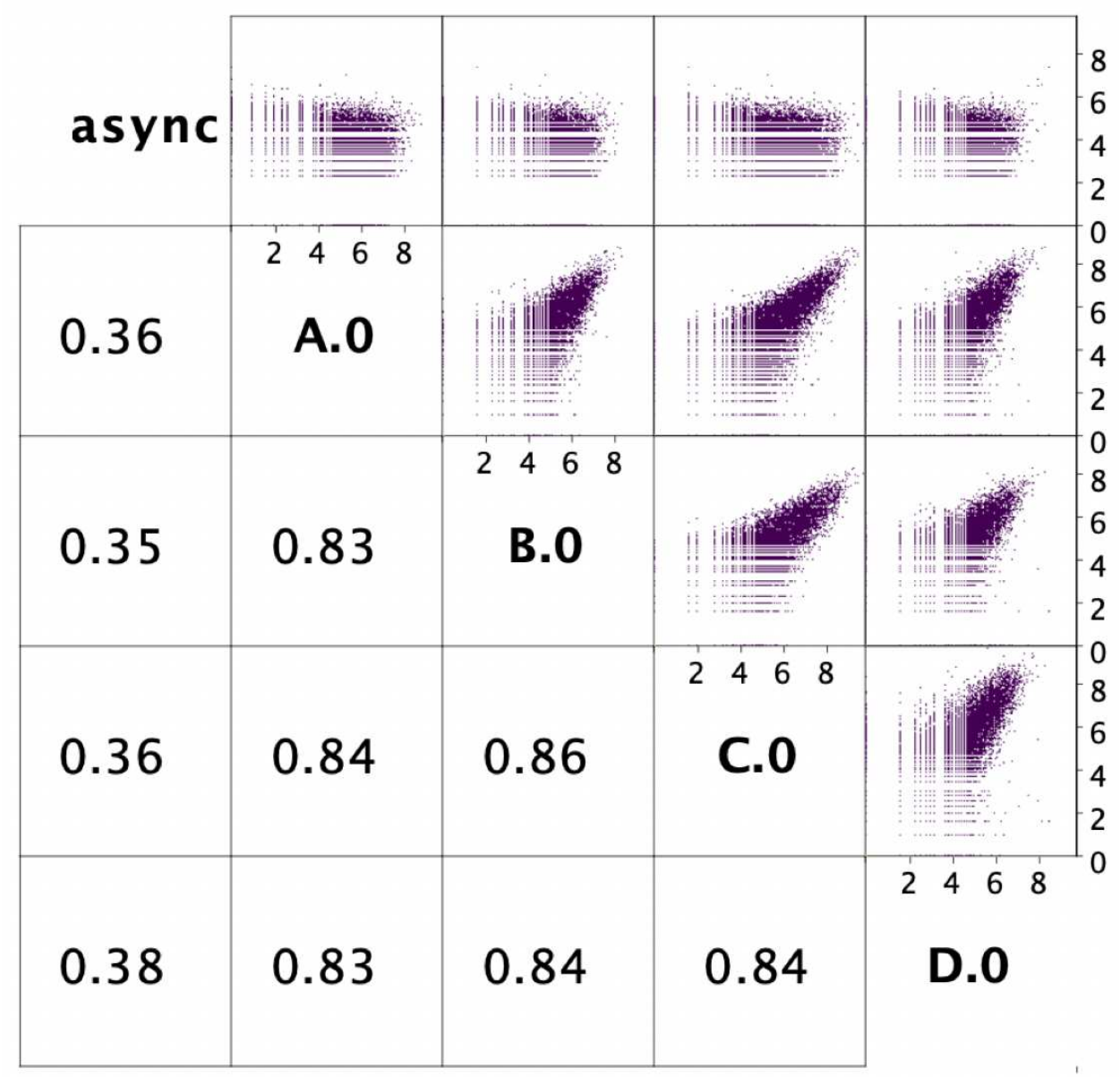


Fig S1.G The correlation of labeling between the four biological replicates (A.0, B.0, C.0, D.0) of the 0-minute dataset and one asynchronous dataset. The correlations between the biological replicates are higher than those between the technical replicates because the biological replicates are larger, reducing the counting noise in infrequently labeled regions of the genome.

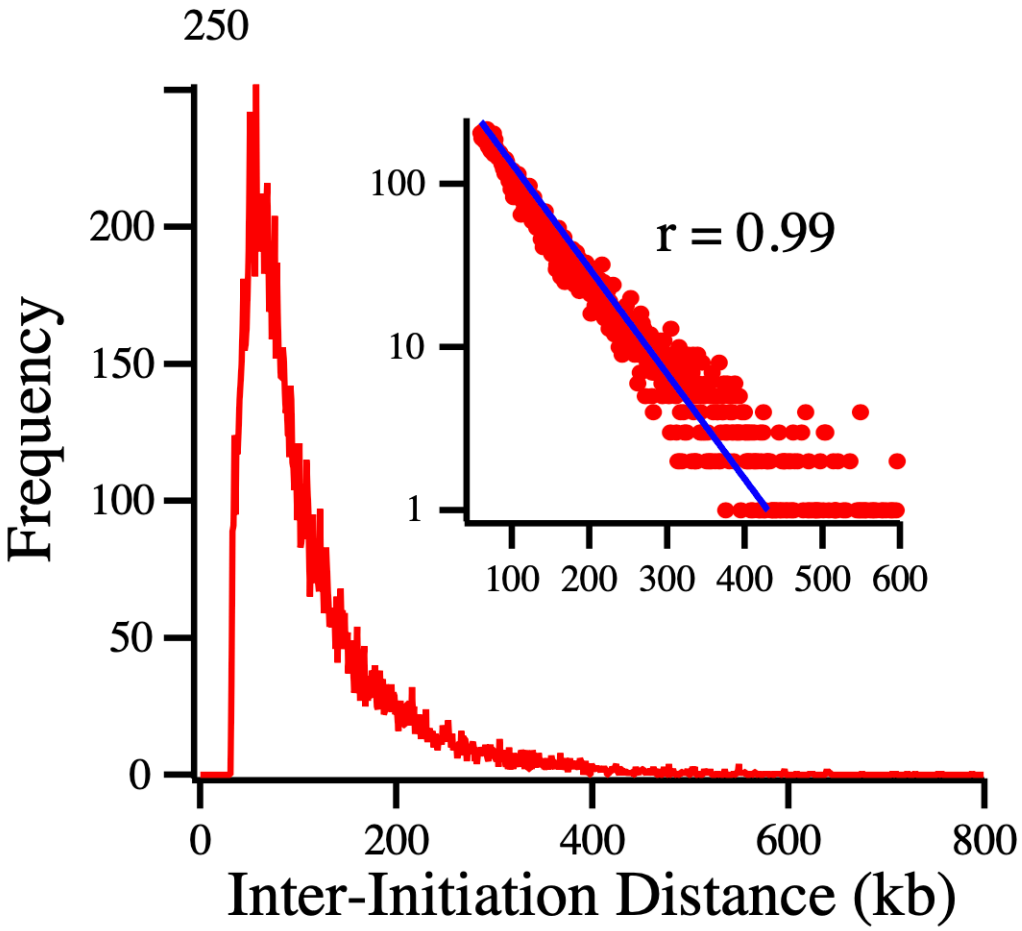


Fig S2.A The distribution of the inter-initiation distances in the combined 0' dataset measured as the distance from the middle of neighboring replications tracks for fibers that have multiple tracks. The average of the distribution is 111 kb and the mode of distribution is 57 kb. Inset: The distribution, plotted on a log y-axis from 60 to 600 kb, fit to an exponential curve (r = 0.99). The exponential distribution of the inter-replication-track distances indicates that the distribution of initiation events on this length scale is random.

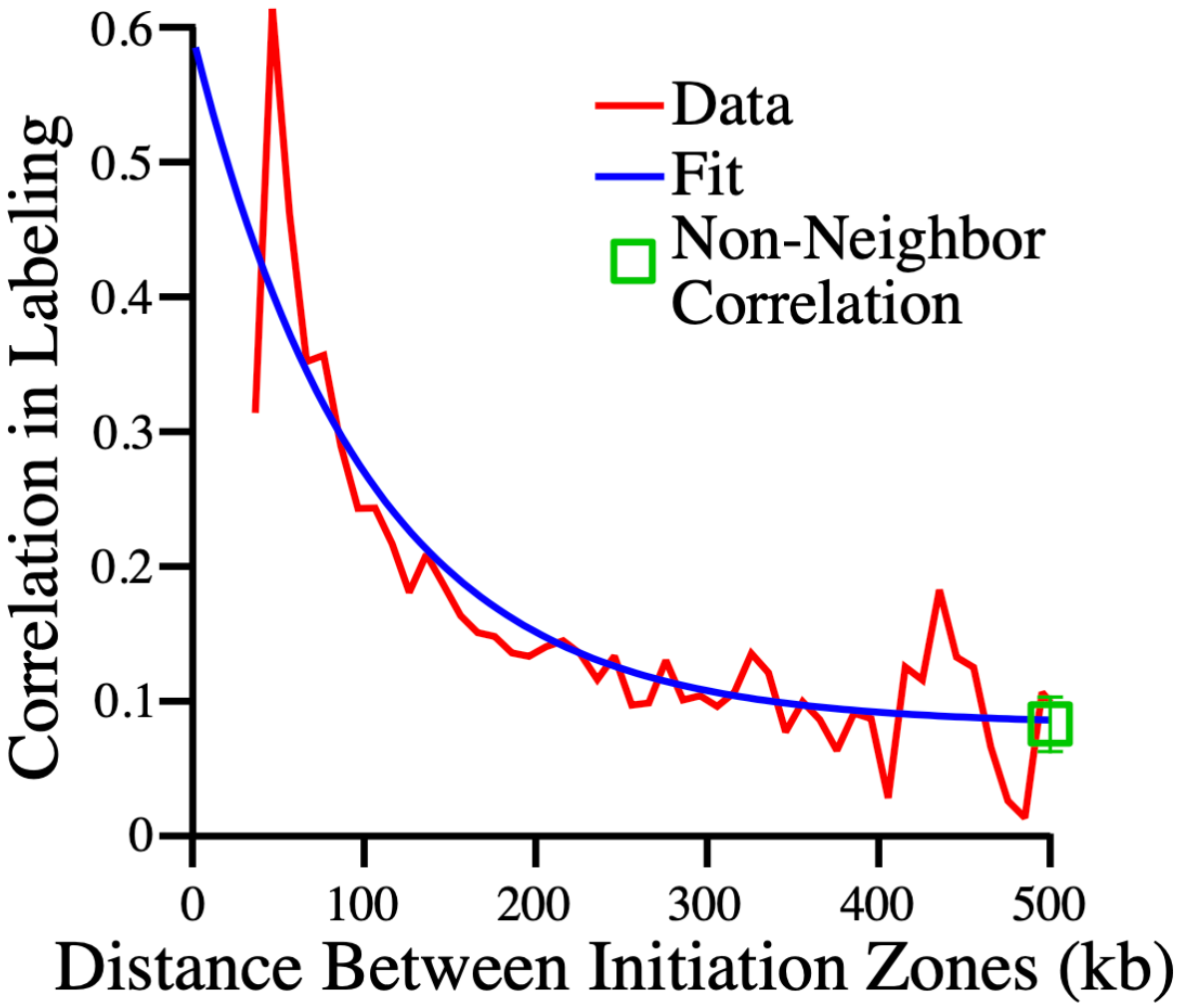


Fig S2.B The correlation between the probability of replication at neighboring IZs. The non-neighbor correlation (0.06±0.01, mean±s.e.m.) is the average correlation between any two IZs on one fiber more than 200 kb apart, irrespective of the number of intervening IZs.

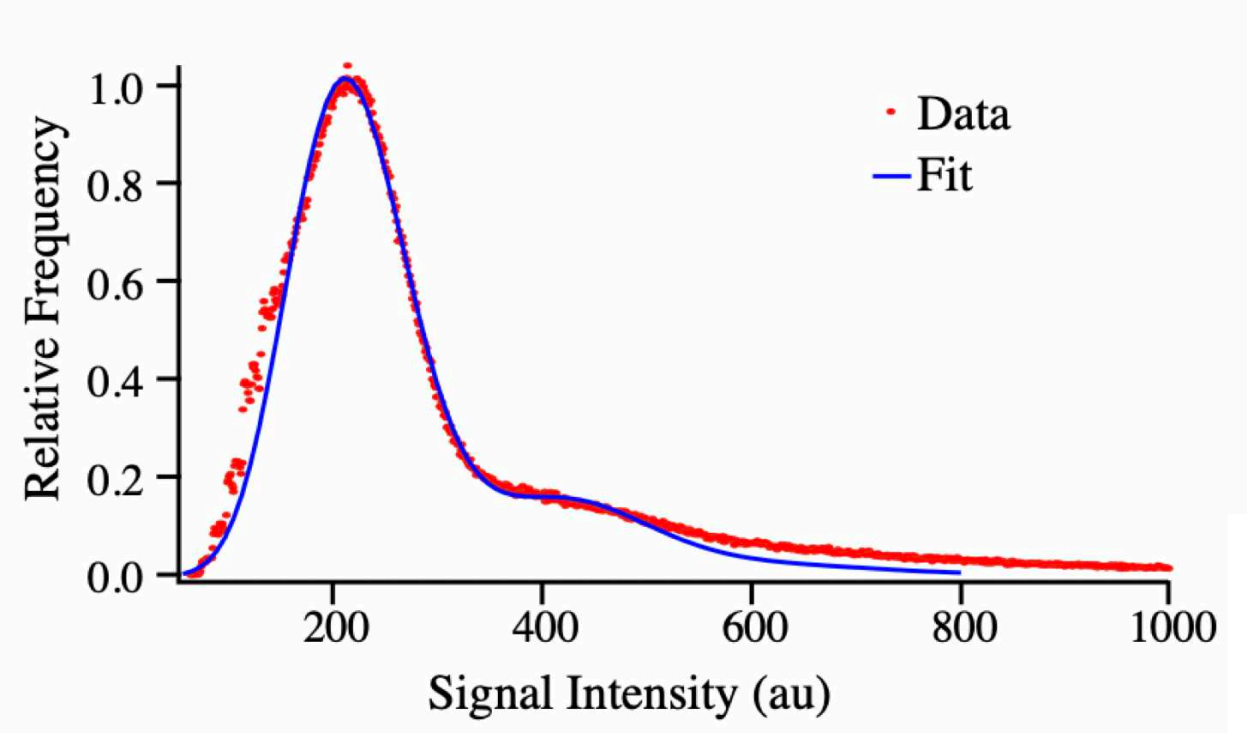


Fig S3.A The distribution of the intensities of the incorporated fluorescence signals in the D.0 0-minute dataset. The fit (obtained from the first four terms of Eq. 5 of Supplemental Mathematical Methods) predicts that about 80% of observed signals are single fluorophores and that the other 20% are multiple fluorophores sufficiently close together that they are not resolved by the Saphyr optics. This estimate of 80% single fluorophores is consistent with an average inter-signal distance of 4 kb and 1.3 kb resolution of the Saphyr, both of which parameters can be inferred from the distribution of inter-signal distances. The distribution of intensities in the other datasets are similar, although differences in the Saphyr optical calibration on different runs introduces variation into the absolute value of the measured fluorescent intensities.

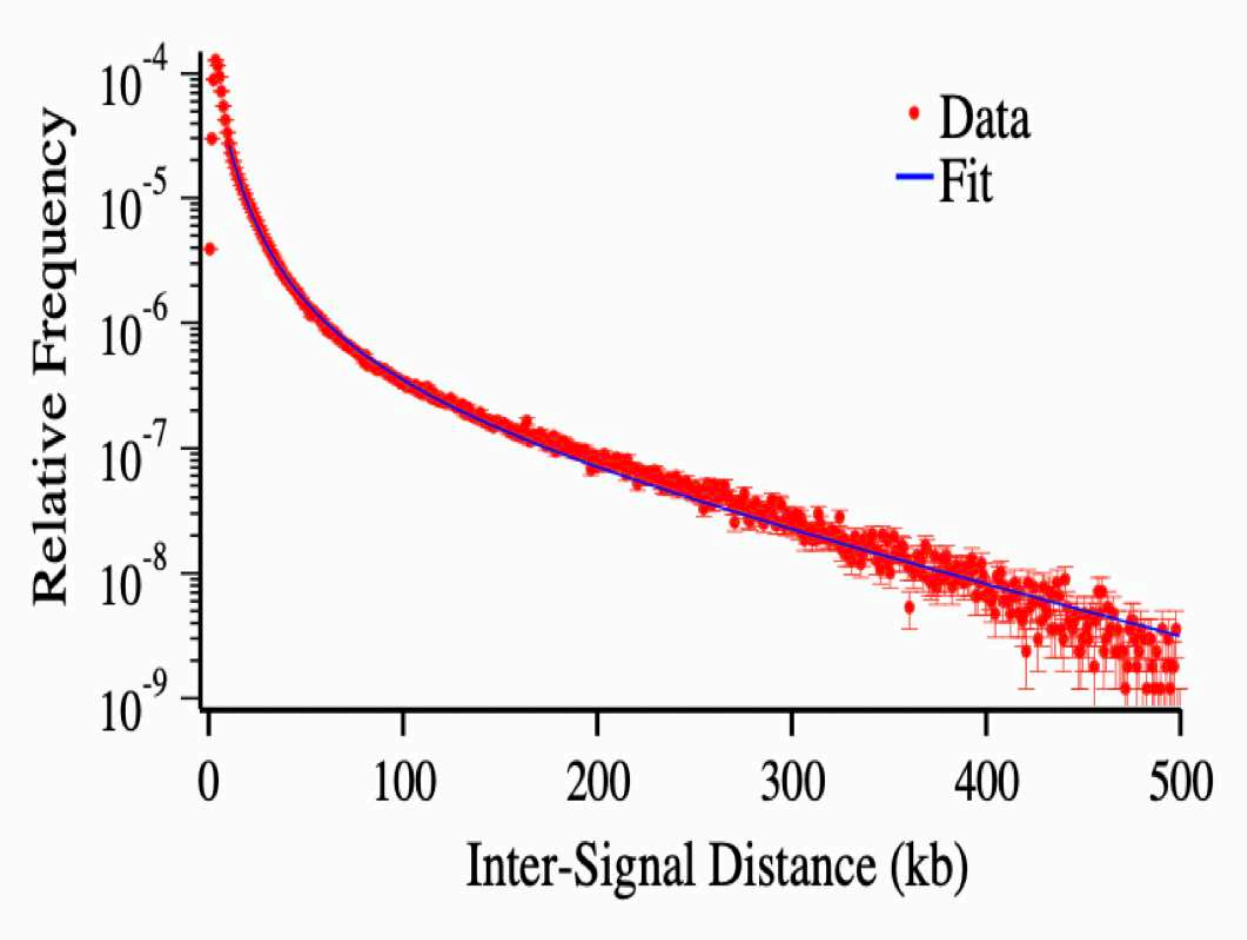


Fig S3.B The distribution of inter-signal distances in the combined 0-minute dataset. The fit to the data (Eq. 14 of Supplemental Mathematical Methods) between 10 and 500 kb predicts an initial labeling frequency of 1 in every 877±17 thymidines and a depletion half-length of 74.5±0.7 kb. Similar fits for the asynchronous HeLa and H9 datasets predict labeling frequencies of 1/1025 and 1/850 and depletion half lengths of 57 and 48 kb, respectively. The similar labeling densities suggest the nucleotide uptake is similar in all three experiments, whereas the shorter depletion half-length is consistent with previous reports that the number of forks increases during S phase, which would consume nucleotides more quickly (Yang and Bechhoefer, 2008; Goldar et al., 2009).

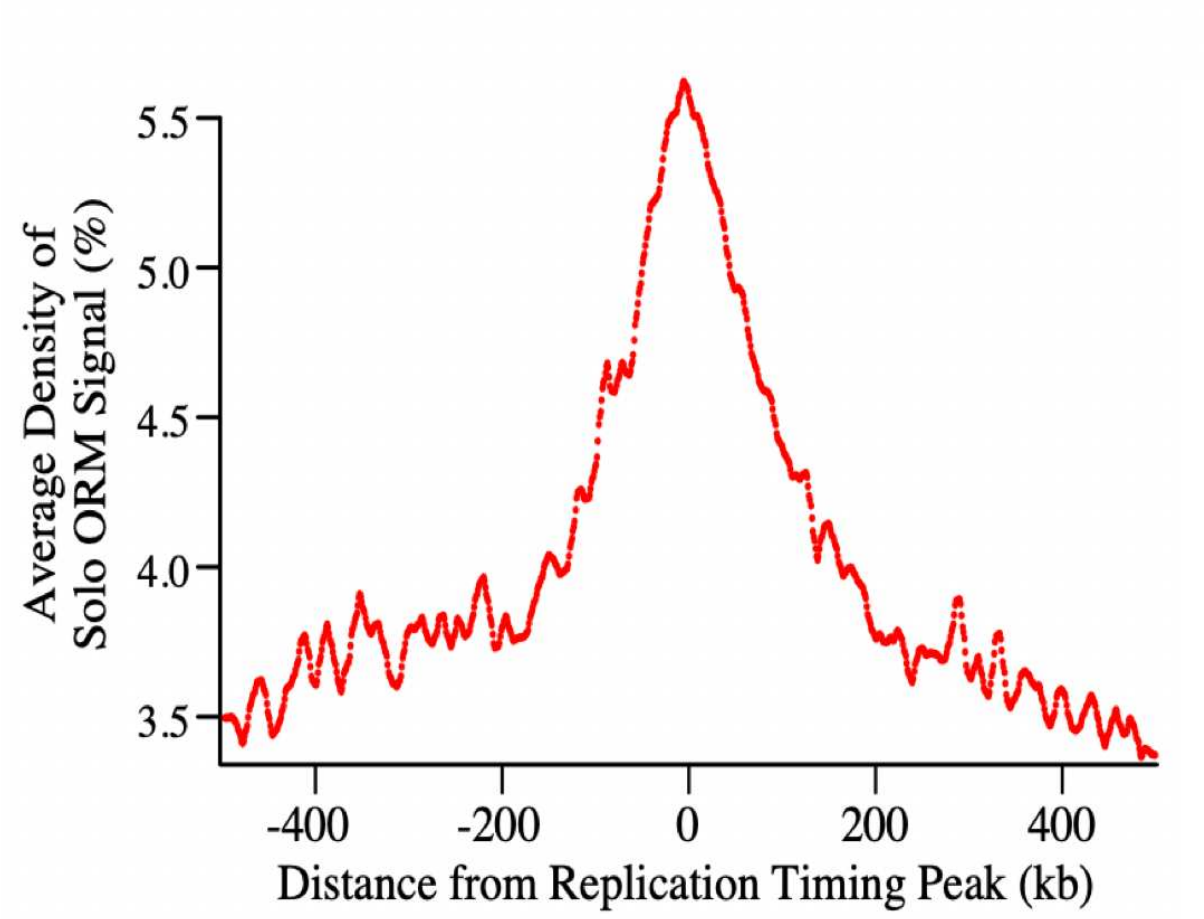


Fig S3.C The enrichment of ORM signals in solo-signal replication tracks from the combined 0-minute dataset around early replication-timing peaks, those that replicate in the first quarter of S phase.

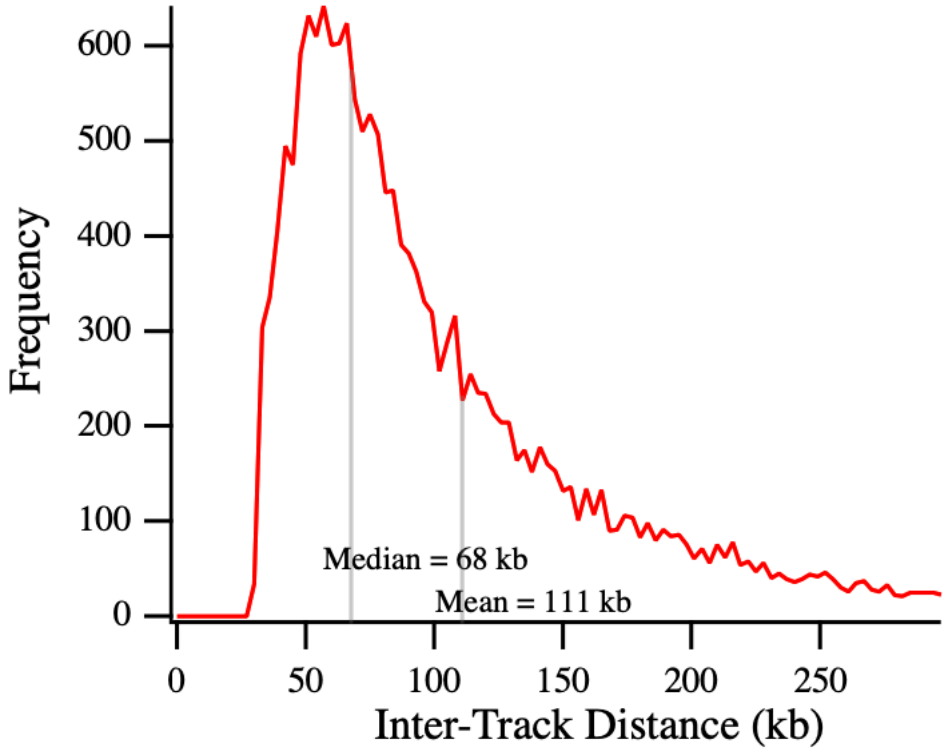


Fig S3.D The distribution of inter-replication-track lengths.

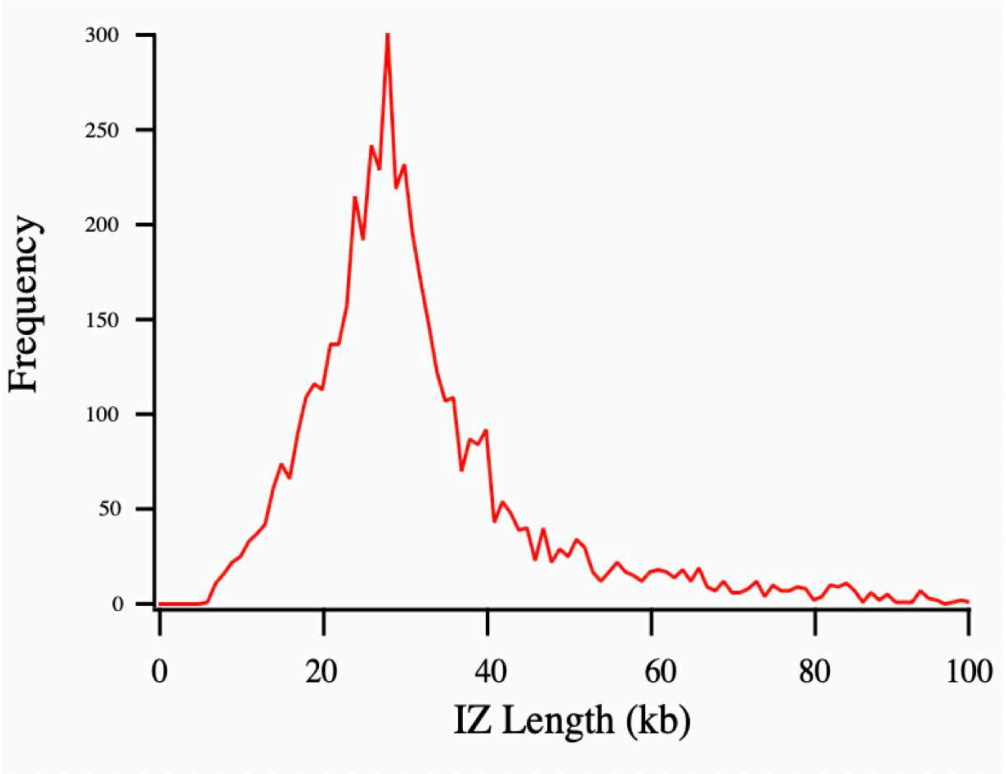


Fig S3.E The distribution of IZ lengths.

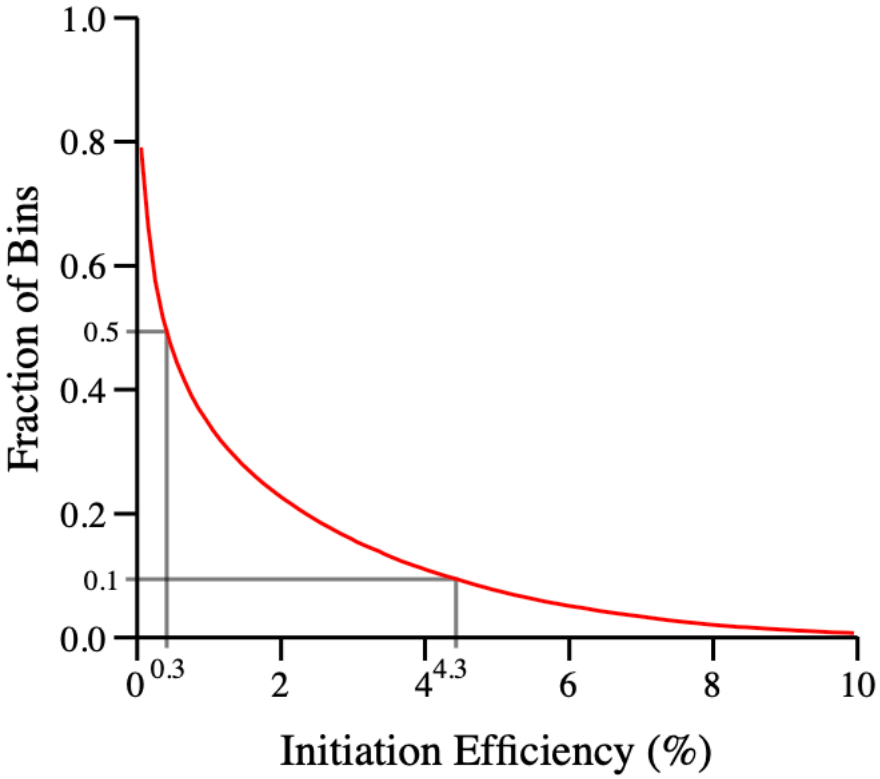


Fig S3.F The fraction of 50 kb genomic bin with initiation efficiency greater than indicated on the x axis. 50% of bins have an initiation efficiency greater that 0.3% and 10% of bins have an initiation efficiency of greater that 4.3%.

Fig S4 The ORM Genome Browser allows interactive visualization the HeLa ORM data. Shown is a screen shot of the Fibers track of synchronous data. It shows fibers as gray bars, ORM signals as yellow hash marks, and inferred replication tracks as black lines. Only fibers labeled with ORM signals are displayed because only~5% of fibers are labeled; displaying all fibers is impractical. The browser can also display only the replication tracks, to provide a more easily- visualized view of replication initiation. It can also show the fibers and the replication tracks from the asynchronous data. Two low-efficiency initiation zones are shown, 2353 (3%) and 2354 (2%), because higher-efficiency initiation zones contain too many labeled fibers. Note that, although the signals and replication tracks are concentrated around the initiation zones, many lay outside of them and none are concentrated in discrete areas.

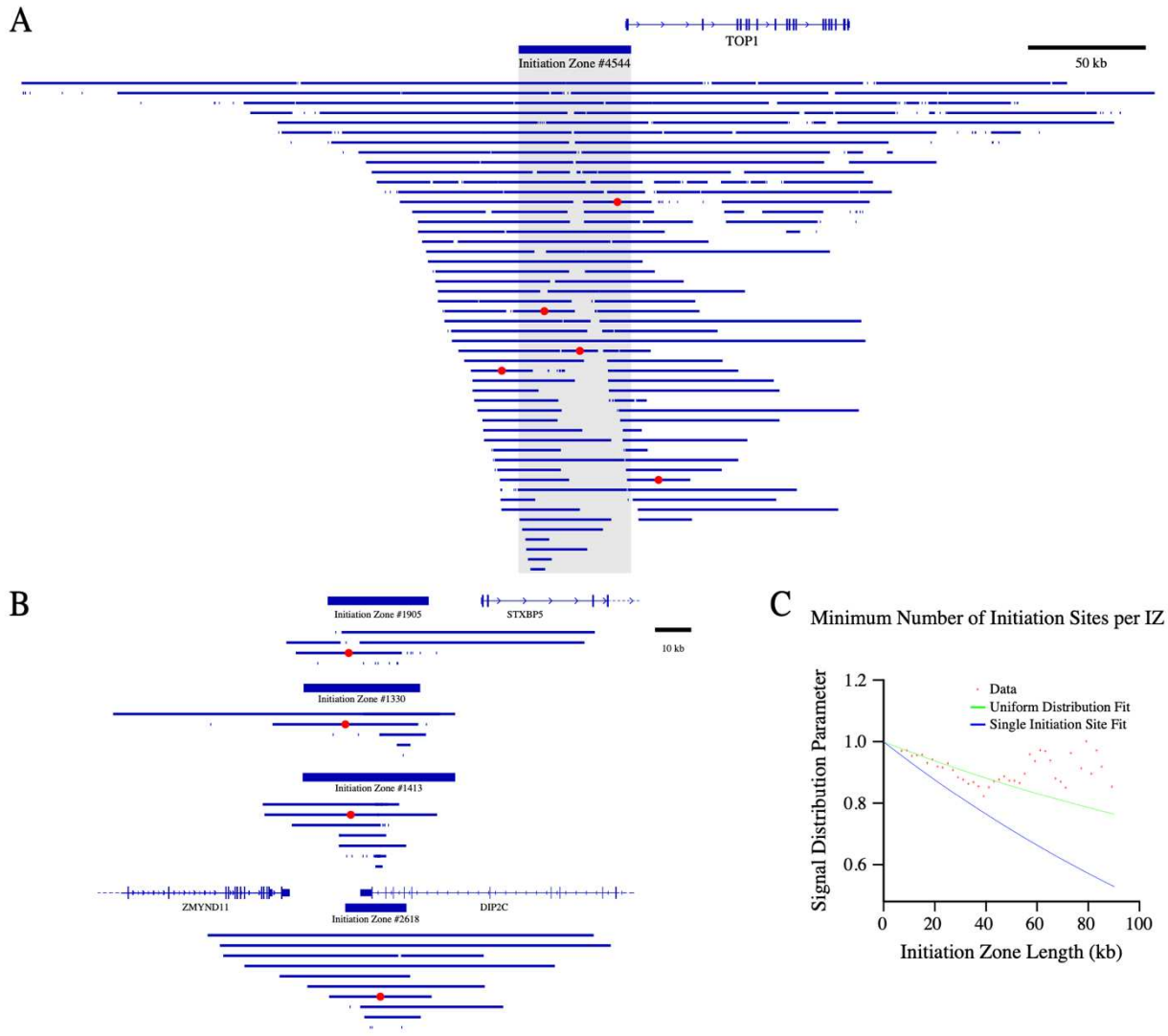


Fig S5.A-C Distribution of Replication Tracks within Initiation Zones.

A) The distribution of replication tracks in the merged 0 minute dataset at the Top1 locus. The Top1 IZ has an estimated minimum of five initiation sites because the five replication track centers indicated in red are all 15 kb away from each other.

B) The distribution of replication tracks at four examples of IZs for which our estimate of the minimum number of initiation sites is 1.

C) The distribution of signal across IZs. The ratio of signal frequency at the IZ center to the IZ boundary is plotted versus IZ length. This value is expected to decrease more quickly in IZs that predominantly have a single initiation site (Eq. 25 of Supplemental Mathematical Methods) than if initiation is distributed across the IZ (Eq. 28 of Supplemental Mathematical Methods). The distribution across IZs shorter than 55 kb is consistent with a uniform distribution of initiation sites. At longer lengths, we actually see more signal at the edges of the IZs than it the center. One possible explanation for this phenomenon is that larger IZs may actually be two smaller IZs fused together such that there is more initiation sites towards the edges of the fused IZ and less initiation in the center, which is actually between the two constituent IZs.

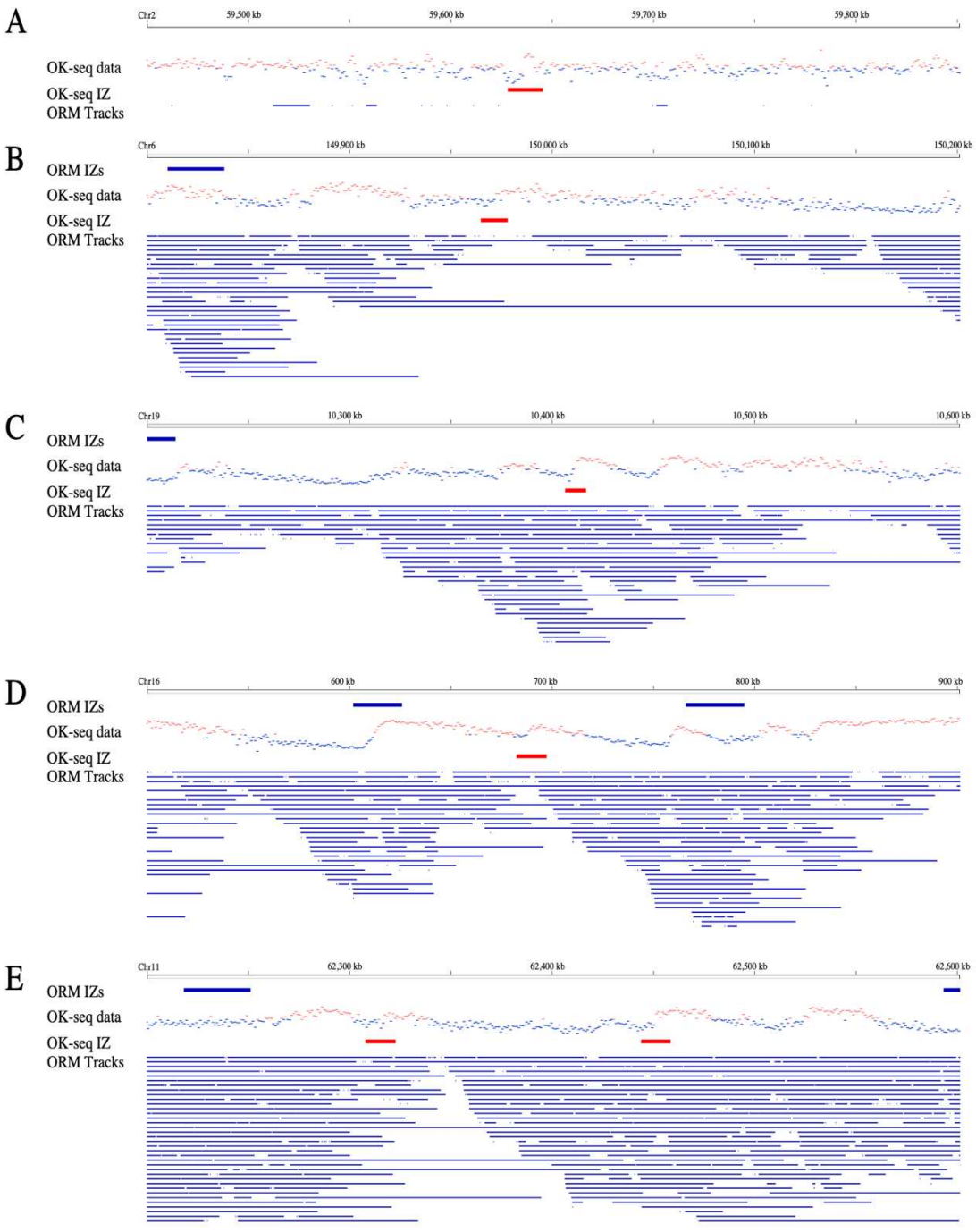


Fig S6.A-E Reanalysis of Potentially Discrete OK-Seq Initiation Zones.

We reexamined the 66 OK-seq IZs that were reported to be less that 5 kb wide (Petryk et al., 2016).

A, **B**) 53 are in regions of noisy OK-seq data. Of those, 24 are in late-replicating regions and appear to be in regions with extensive bi-direction replication. Panel A is an example of one such zone. However, since there is little ORM data in these regions, we can say little more about them. 29 are in early replicating regions, but none of them correlate with numerous ORM segments. Panel B is an example of one such zone. We conclude that they are not active IZs in our ORM data and probably not active IZs in the OK-seq data, either.

C) 6 are robust transitions that correlate with numerous ORM segments. We conclude that they are IZs in both the OK-seq and ORM data. However, they show broadly dispersed ORM segments, therefore we do not believe they are unusually constrained IZs. Instead, we conclude that they are outliers in the OK-seq data that were identified as unusually narrow due to experimental variation. Panel C is an example of one such zone.
D, E) 7 are robust transitions that do not correlate with numerous ORM segments. They could be IZs present in the OK-seq HeLa cell line but absent in the ORM HeLa cell line. Alternatively, they could be translocation break points in the OK-seq HeLa cell line relative to the hg19 reference sequence. Such breakpoints would explain both the sharpness of the transition and the absence of these putative IZs from the ORM data. Panel D is an example of one such zone. Panel E show two such zones that can be explained by an inversion between them.

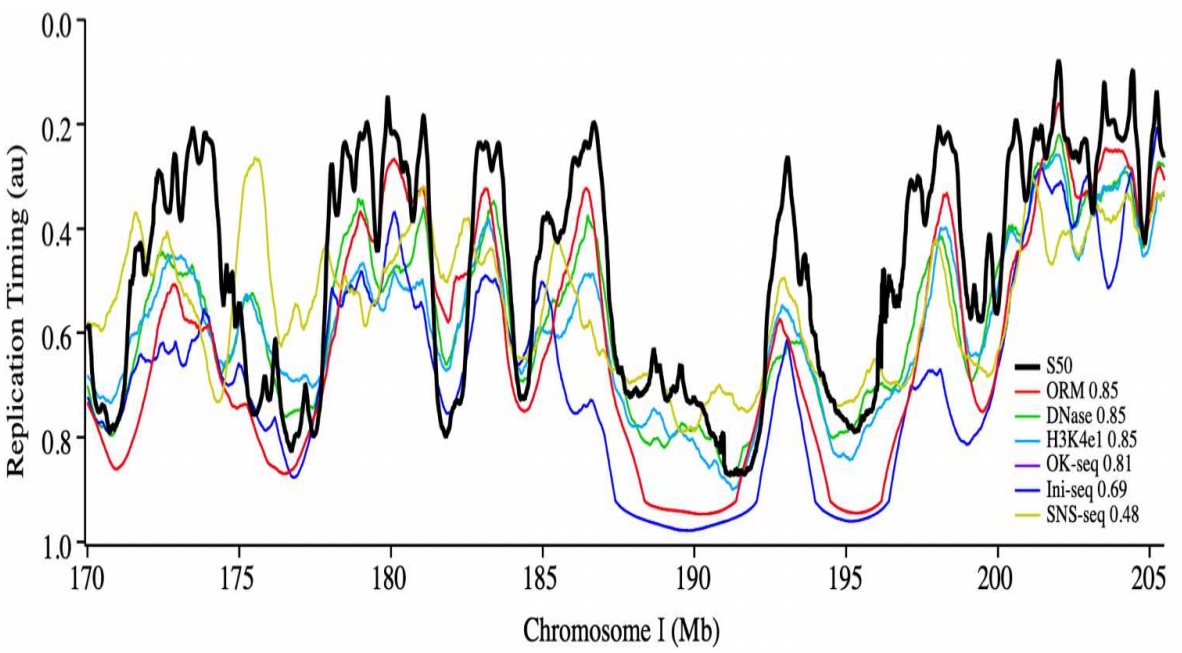


Figure S7 Simulated Replication Timing Profiles.
Comparison between experimentally determined HeLa replication timing (S50) and replication timing predicted from ORM, DNase I hypersensitivity (Bernstein et al., 2012), OK-Seq (Petryk et al., 2016), ini-seq (Langley et al., 2016) and SNS-seq (Picard et al., 2014) data using a stochastic model (Gindin et al., 2014a). The Spearman correlation coefficients with replication timing are shown in the legend.

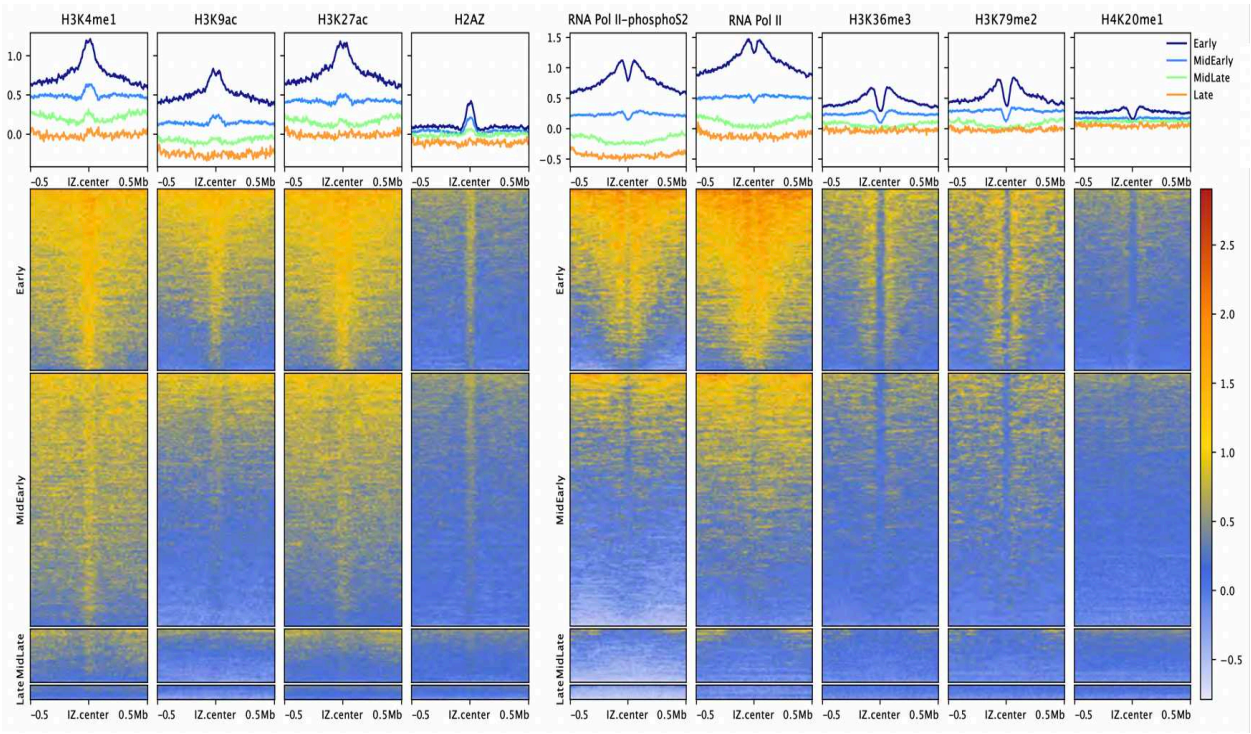


Figure S8.A The enrichment of GC content relative to ORM IZs. The upper panels show the average % GC content signal around all IZs. The lower panels show heat maps of the % GC content at each IZ.

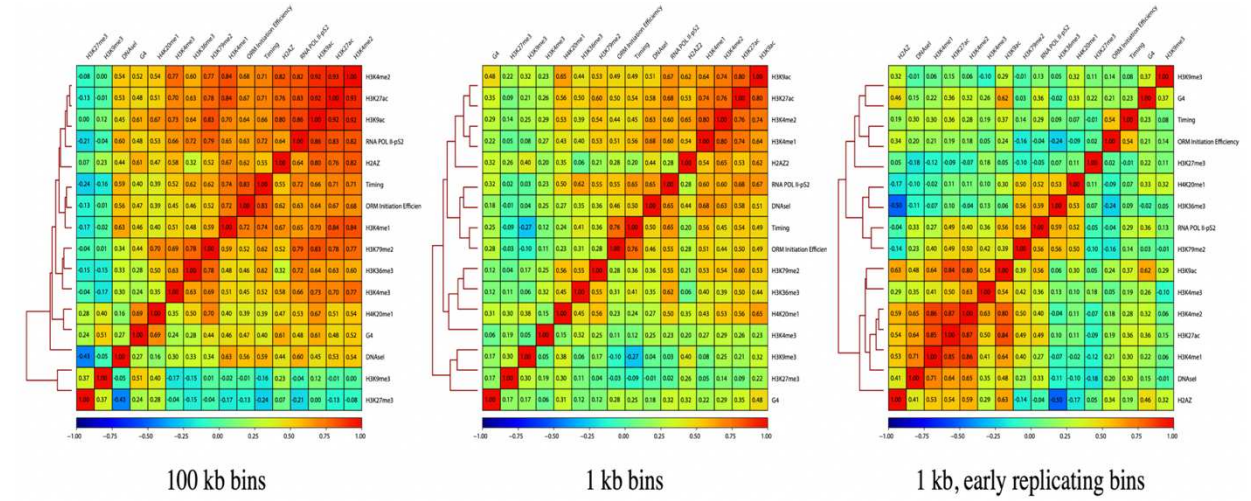


Figure S8.B Correlation heat maps at various resolution. The left panel shows 100 kb resolution, which does not resolve enhancers, promoters and transcription units. Therefore, features associated with all three correlates with ORM signal. The center panel shows 1 kb resolution, which resolves enhancers, promoters and transcription units. However, the correlation is dominated by replication timing, creating a correlation between ORM IZs and transcription units, which both tend to replicate early. The right panel shows 1 kb resolution for the earliest-replicating quarter of the genome. Here, enhancer-enriched features, such as H3K4me1, H3K9ac and H3K27ac hypersensitivity, are most strongly correlated, while promoter-enriched features, such as RNA Pol II, H3K4me3, and are more weakly correlated and elongation-enriched features, such as H4K20me1, H3K79me2 and H3K36me3, are anti-correlated.

RÉSUMÉ

La réplication de l'ADN est régulée par l'emplacement et le moment de l'initiation de la réplication. Par conséquent, beaucoup d'efforts ont été investis dans l'identification et l'analyse des sites d'initiation de la réplication dans les cellules humaines. Cependant, la nature hétérogène de la cinétique de réplication eucaryote et la faible efficacité de l'utilisation du site d'initiation individuelle chez les métazoaires a rendu difficile la cartographie de l'emplacement et du moment de l'initiation de la réplication dans les cellules humaines. Une solution potentielle au problème de la cartographie de la réplication humaine est l'analyse dans les molécules uniques. Cependant, les approches actuelles ne fournissent pas le débit requis pour les expériences à l'échelle du génome humaine. Pour relever ce défi, nous avons développé la cartographie de réplication optique (Optical Replicaiton Mapping - ORM), une approche de molécule unique à haut débit pour cartographier l'ADN nouvellement répliqué, et l'avons utilisée pour cartographier les événements d'initiation précoce dans les cellules humaines. La nature de molécule unique de nos données, et une couverture totale de plus de 2000 fois du génome humain sur 27 millions de fibres d'une longueur moyenne d'environ 300 kb, nous permettent d'identifier les sites d'initiation et leur probabilité d'initiation avec une grande confiance. En particulier, pour la première fois, nous sommes en mesure de mesurer à l'échelle du génome humain l'efficacité absolue de l'initiation de la réplication. Nous constatons que la distribution de l'initiation de la réplication humaine est cohérente avec l'initiation inefficace et stochastique de complexes d'initiation potentiels distribués de manière hétérogène enrichis en chromatine accessible. En particulier, nous constatons que les sites d'initiation de la réplication humaine ne sont pas limités à des origines de réplication bien définies, mais sont plutôt répartis sur de larges zones d'initiation constituées de nombreux sites d'initiation. De plus, nous ne trouvons aucune corrélation des événements d'initiation entre les zones d'initiation voisines. Bien que la plupart des événements d'initiation précoce se produisent dans les régions à réplication précoce du génome, un nombre significatif se produit dans les régions tardives. Le fait que les sites d'initiation dans les régions tardive aient une certaine probabilité d'initiation au début de la phase S suggère que la principale différence entre les événements d'initiation dans les régions à réplication précoce et tardive est leur probabilité intrinsèque d'initiation, et n'est pas due à une différence qualitative dans leur distribution de temps d'initiation. De plus, la modélisation de la cinétique de réplication démontre que la mesure de l'efficacité d'initiation de la zone d'initiation au début de la phase S suffit pour prédire le temps d'initiation moyen de ces zones tout au long de la phase S, ce qui suggère en outre que les différences entre les temps d'initiation des zones d'initiation précoce et tardive sont quantitatives plutôt que qualitatives. Ces observations sont cohérentes avec les modèles stochastiques de la régulation de l'initiation et suggèrent que la régulation stochastique de la cinétique de réplication est une caractéristique fondamentale de la réplication chez eucaryotes, conservée de la levure à l'homme.

KEYWORDS

Réplication d'ADN, Origine de la Réplication, Molécule Unique, Cartographie Optique, Timing de la Réplication

ABSTRACT

DNA replication is regulated by the location and timing of replication initiation. Therefore, much effort has been invested in identifying and analyzing the sites of human replication initiation. However, the heterogeneous nature of eukaryotic replication kinetics and the low efficiency of individual initiation site utilization in metazoans has made mapping the location and timing of replication initiation in human cells difficult. A potential solution to the problem of human replication mapping is single-molecule analysis. However, current approaches do not provide the throughput required for genome-wide experiments. To address this challenge, we have developed Optical Replication Mapping (ORM), a high-throughput single-molecule approach to map newly replicated DNA and used it to map early initiation events in human cells. The single-molecule nature of our data, and a total of more than 2000-fold coverage of the human genome on 27 million fibers averaging ~300 kb in length, allow us to identify initiation sites and their firing probability with high confidence. In particular, for the first time, we are able to measure genome-wide the absolute efficiency of human replication initiation. We find that the distribution of human replication initiation is consistent with inefficient, stochastic initiation of heterogeneously distributed potential initiation complexes enriched in accessible chromatin. In particular, we find sites of human replication initiation are not confined to well-defined replication origins but are instead distributed across broad initiation zones consisting of many initiation sites. Furthermore, we find no correlation of initiation events between neighboring initiation zones. Although most early initiation events occur in early-replicating regions of the genome, a significant number occur in late replicating regions. The fact that initiation sites in typically late-replicating regions. The fact that initiation sites in typically late-replicating regions have some probability of firing in early S phase suggests that the major difference between initiation events in early and late replicating regions is their intrinsic probability of firing, as opposed to a qualitative difference in their firing-time distributions. Moreover, modeling of replication kinetics demonstrates that measuring the efficiency of initiation-zone firing in early S phase suffices to predict the average firing time of such initiation zones throughout S phase, further suggesting that the differences between the firing times of early and late initiation zones are quantitative, rather than qualitative. These observations are consistent with stochastic models of initiation-timing regulation and suggest that stochastic regulation of replication kinetics is a fundamental feature of eukaryotic replication, conserved from yeast to humans.

KEYWORDS

DNA replication, Replication Origin, Single-Molecule, Optical Mapping, Replication Timing