2020-08-24

Genome-Wide Mapping of Human DNA Replication by Optical Replication Mapping Supports a Stochastic Model of Eukaryotic Replication Timing [preprint]

Weitao Wang
Paris Sciences et Lettres University

Et al.

Repository Citation

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

Weitao Wang[1]*, Kyle Klein[2]*, Karel Proesmans[3]*, Hongbo Yang[4]*, Claire Marchal[2§], Xiaopeng Zhu[5], Tyler Borman[6], Alex Hastie[7], Zhiping Weng[6], John Bechhoefer[3†], Chun-Long Chen[1,8†], David M. Gilbert[2†] and Nicholas Rhind[9†]

1) Institut Curie, PSL Research University, CNRS UMR 3244, Paris, France
2) Florida State University, Department of Biological Science, Tallahassee, FL, USA
3) Simon Fraser University, Department of Physics, Burnaby, British Columbia, Canada
4) Northwestern University, Feinberg School of Medicine, Chicago, Department of Biochemistry and Molecular Genetics, Illinois, USA
5) Carnegie Mellon University, Computational Biology Department, Pittsburgh, PA, USA
6) University of Massachusetts Medical School, Program in Bioinformatics and Integrated Biology, Worcester, MA, USA
7) Bionano Genomics, San Diego, CA, USA
8) Sorbonne University, Paris, France
9) University of Massachusetts Medical School, Department of Biochemistry and Molecular Pharmacology, Worcester, MA, USA
*) These authors made comparable contributions.
§) Present address: Neurobiology, Neurodegeneration & Repair Laboratory, National Eye Institute, National Institutes of Health, Bethesda, MD, USA
†) Corresponding authors: johnb@sfu.ca, chunlong.chen@curie.fr, gilbert@bio.fsu.edu, nick.rhind@umassmed.edu

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 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.

Introduction

The human genome is replicated from approximately 50,000 distinct initiation events in each cell cycle (Masai et al., 2010). However, identifying the location and firing timing of human replication origins is an ongoing challenge (Hyrien, 2015). A major part of this challenge is the inefficient and heterogeneous nature of mammalian origin firing (Berezney et al., 2000; Taylor, 1977). Even in cases where mammalian origins are thought to occur at relatively well-defined sites, they appear to fire with as little as 10% efficiency (Dijkwel et al., 2002); in the more common case where initiation appears to occur within broad initiation zones, spanning tens of kilobases (Hamlin et al., 2008), they fire with less than 1% efficiency (Demczuk et al., 2012). Therefore, population-based bulk approaches to mapping origins, which average origin-firing behavior across large populations of cells, must contend with low signal-to-noise ratios. As a result, although a number of bulk approaches have been used to map origins in the human genome (Mesner et al., 2013; Besnard et al., 2012; Cayrou et al., 2011; Foulk et al., 2015; Langley et al., 2016; Macheret and Halazonetis, 2018; Dellino et al., 2013; Petryk et al., 2016), there is little concordance among them (Langley et al., 2016; Mesner et al., 2013).

The heterogeneous nature of mammalian origins may be influenced by the non-site-specific nature of the mammalian origin recognition complex (ORC) DNA binding. Unlike budding yeast ORC, which binds a specific 17-bp sequence (Theis and Newlon, 1997; Eaton et al., 2010), mammalian ORC shows only a weak affinity for AT-rich sequences *in vitro* (Vashee et al., 2003; De Carli et al., 2018) and no discernible sequence preference *in vivo* (Miotto et al., 2016). G-quadruplex (G4) sequences have been proposed to function as origins (Cayrou et al., 2015; Valton et al., 2014; Prorok et al., 2019), but this hypothesis is controversial (Foulk et al., 2015; Miotto et al., 2016). A simple explanation for the observed patterns of ORC binding and replication initiation is that ORC binds non-specifically to accessible, nucleosome-free DNA (Lubelsky et al., 2011). In particular, ORC binding and replication initiation correlate well with DNase I accessibility across metazoan genomes (Miotto et al., 2016; Gindin et al., 2014b). Furthermore, MCMs can slide after loading, further dispersing sites of initiation (Gros et al., 2015; Edwards et al., 2002; Powell et al., 2015). For these reasons, we will refrain from referring to mammalian sites of replication initiation as origins but rather refer to specific instances of replication initiation as *initiation events* occurring at *initiation sites* and to areas enriched in initiation sites as *initiation zones*.

Although the specific locations of replication initiation are unclear, the general characteristics of mammalian sites of initiation have been described. Initiation is enriched in euchromatic promoter and enhancer regions, consistent with its correlation with accessible chromatin (Ganier et al., 2019; Petryk et al., 2016; Pourkarimi et al., 2016; Cayrou et al., 2015). Although initiation tends to be enriched in euchromatic regions, initiation must also occur in heterochromatin. Initiation in heterochromatin appears to be even more heterogeneous, making heterochromatic initiation sites even less well understood (Petryk et al., 2016; Cayrou et al., 2015).

Much effort has also been put into understanding the relative timing of DNA replication across the genome (Rhind and Gilbert, 2013; Fragkos et al., 2015; Marchal et al., 2019). The replication timing of a genomic region is primarily determined by the timing of initiation within that region. Based on population replication timing data, metazoan genomes can be divided into regions of similar local replication timing referred to as constant timing regions (Hiratani et al., 2008; Farkash-Amar et al., 2008; Hansen et al., 2010). Constant timing regions consist of smaller units whose replication timing can be regulated during development; these smaller units are called replication domains (Hiratani et al., 2008). Replication domains tend to align with topologically associated domains (TADs) and are thought to represent discrete functional units of the genome (Pope et al., 2014). The replication timing of replication domains correlates with the level of expression of the genes they contain and their chromatin modifications. However, whether the replication timing is a cause, a consequence, or an independent correlate of chromatin structure is a matter of debate (Rhind and Gilbert, 2013).

In budding yeast, where replication origin locations are well defined, origin firing is stochastic, with each origin firing with a specific probability, independent of neighboring origins (Czajkowsky et al., 2008; Yang et al., 2010; de Moura et al., 2010). Such stochastic firing leads to reproducible replication profiles at the population level, because more efficient origins are more likely to fire early and therefore, on average, have early replication times; by contrast, inefficient origins usually fire late or are passively replicated (Rhind et al., 2010). The heterogeneous and inefficient nature of metazoan replication initiation is also consistent with stochastic initiation. Simulations with a stochastic firing model, in which initiation is regulated only by a local-firing-probability function, faithfully reproduce experimental genome-wide replication timing profiles, suggesting that no deterministic timing program is required (Gindin et al., 2014b). Furthermore, if the local initiation rate is predicted by DNase I hypersensitivity, the simulation closely matches experimental results, consistent with the observed correlation between promoters and enhancers, which are DNase I hypersensitive, and initiation frequency (Gindin et al., 2014b). On the other hand, the reproducible replication timing of individual replication domains measured in single cells has led to the suggestion that replication within those domains initiates at defined times in most cells in the population (Dileep and Gilbert, 2018). Furthermore, neighboring initiation sites have been proposed to show both cooperative firing and lateral inhibition (Cayrou et al., 2011; Guilbaud et al., 2011; Löb et al., 2016), neither of which are consistent with strictly stochastic models. Therefore, 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).

A powerful solution to the problems of heterogeneity and low signal-to-noise ratios is single-molecule analysis, which allows the identification of sites of replication initiation on individual DNA fibers (Técher et al., 2013). Traditional single-molecule approaches—such as fiber autoradiography (Huberman and Riggs, 1968), DNA combing (Herrick and Bensimon, 1999; Anglana et al., 2003; Kaykov et al., 2016) and SMARD (Norio and Schildkraut, 2001)—have provided critical insight into the location and firing kinetics of mammalian replication origins. However, these techniques are restricted to the analysis of at most a few genomic loci. Nanopore-sequencing-based methods offer the potential to map replication genome-wide at high resolution (Müller et al., 2019; Georgieva et al., 2019; Hennion et al., 2020), but the current throughput of nanopore sequencing does not allow genome-wide analysis of metazoan genomes and current average read lengths are on the order of only 30 kb. Optical mapping technologies provide ample throughput on individual long (150 kb to 2 Mb) DNA fibers (Lam et al., 2012) and can be combined with nucleotide analog incorporation to map DNA replication (De Carli et al., 2018). However, the current approach relies on DNA replication in extracts (De Carli et al., 2018), precluding *in vivo* analysis of replication initiation.

We have developed Optical Replication Mapping (ORM), a single-molecule technique to investigate spatial and temporal distribution of origin firing that combines the Bionano Genomics approach to mapping long individual DNA molecules (Lam et al., 2012) with *in vivo* fluorescent nucleotide pulse-labeling (Wilson et al., 2016; Panning and Gilbert, 2005) to directly visualize sites of replication initiation within human cells. This approach affords us excellent signal-to-noise characteristics and deep, genome-wide coverage, allowing us to identify initiation sites active in as few as 0.1% of human cells. We have used ORM to identify and analyze human sites of DNA replication initiation. We have further applied ORM to asynchronous human cells to obtain cell-type-specific replication profiles.

Results

Optical Replication Mapping of Early-Firing Human Initiation Sites

We mapped sites of early replication in the human genome by ORM, combining *in vivo* labeling of replication initiation with Bionano visualization and mapping of long individual DNA molecules (Lam et al., 2012, Figure 1A and Methods;). HeLa S3 cells were synchronized in mitosis by mitotic shake off and arrested at the beginning of S phase with aphidicolin. This protocol allows early initiation events but prevents elongation of replication forks from these initiation events and also prevents subsequent initiation events by activation of the intra-S-phase checkpoint (Feijoo et al., 2001). Cells were then electroporated with fluorescent ATTO647-dUTP, released from aphidicolin, allowed to recover overnight to complete replication, and harvested for analysis on the Bionano Saphyr platform. As a result, cells incorporate the electroporated fluorescent nucleotides in replication forks that initiated during the aphidicolin arrest and any that are initiated immediately after release. However, the transfected nucleotides are rapidly depleted, preventing incorporation at later initiation sites (Wilson et al., 2016). 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 S1, Figures S1A,B), indicating a high degree of synchrony and <1% of contaminating asynchronous S phase cells.

We collected a dataset of 728 Gb across 2.9 million fibers ranging in length from 150 kb to 1.2 Mb, with an average fiber length of 250 kb (Table S1 and Figure S1C). Coverage of the genome was uniform, with an average coverage of 206 fold and a coefficient of variation of 25%. On these fibers, we identified fluorescent incorporation signals and mapped them across the genome (Figure 1A, see Methods for details). To determine the technical reproducibility of our data, we collected it as three technical replicates. The replicates show high reproducibility, with the variation dominated by counting noise in infrequently labeled regions of the genome (r = 0.81, Figure S1D). To determine the biological reproducibility of our assay, we repeated the experiment three times and compared the signal distribution in the four biologically-independent experiments. The four replicates show high correlation (r = 0.83–0.86, Figure S1E), demonstrating low biological variation in our assay. Therefore, we combined the four experiments into one dataset of 11.8 million fibers ranging from 150 kb to 2.2 Mb in length, averaging 285 kb and constituting 1,109-fold coverage of the human genome (Table S1).

The replication-incorporation signal we detect is punctate in nature and uniform in intensity (Figures 1A and S2A). From these observations, we conclude that these incorporation signals consist of individual fluorescent nucleotides. Furthermore, the distributions of signal intensity (Figure S2A) and inter-signal distances (Figure S2B) are consistent with sparse labeling by a pool of fluorescent nucleotides that is depleted by replication, as expected for a one-time bolus of nucleotides delivered by electroporation. The distribution of the signals is predicted by inefficient, random incorporation of fluorescent nucleotides at an initial rate of about 1/900 thymidines that is depleted with a half length of about 75 kb (Figure S2B). Given the lengths of the replication tracks and the sparse nature of the signal within them, we estimate that we can localize initiation events with a resolution of about 15 kb (14.2±0.1 kb, see Methods for details).

The fluorescent signals in our initial ORM dataset labels are enriched in replication initiation zones, as identified by replication timing peaks (Chen et al., 2011), consistent with our prediction that they represent nucleotide incorporation near sites of replication initiation (Figure 1B). To test whether the incorporation signals truly reflected initiation events, we modified our protocol to allow for a delay, ranging from 5' to 90', after aphidicolin release and before transfection (Table S1). We predicted that this approach would result in a gap between two symmetric populations of incorporation signals surrounding the replication initiation sites. As predicted, when we aggregate replication signals around early initiation zones, we see a time-dependent movement of incorporation tracks away from the

Figure 1

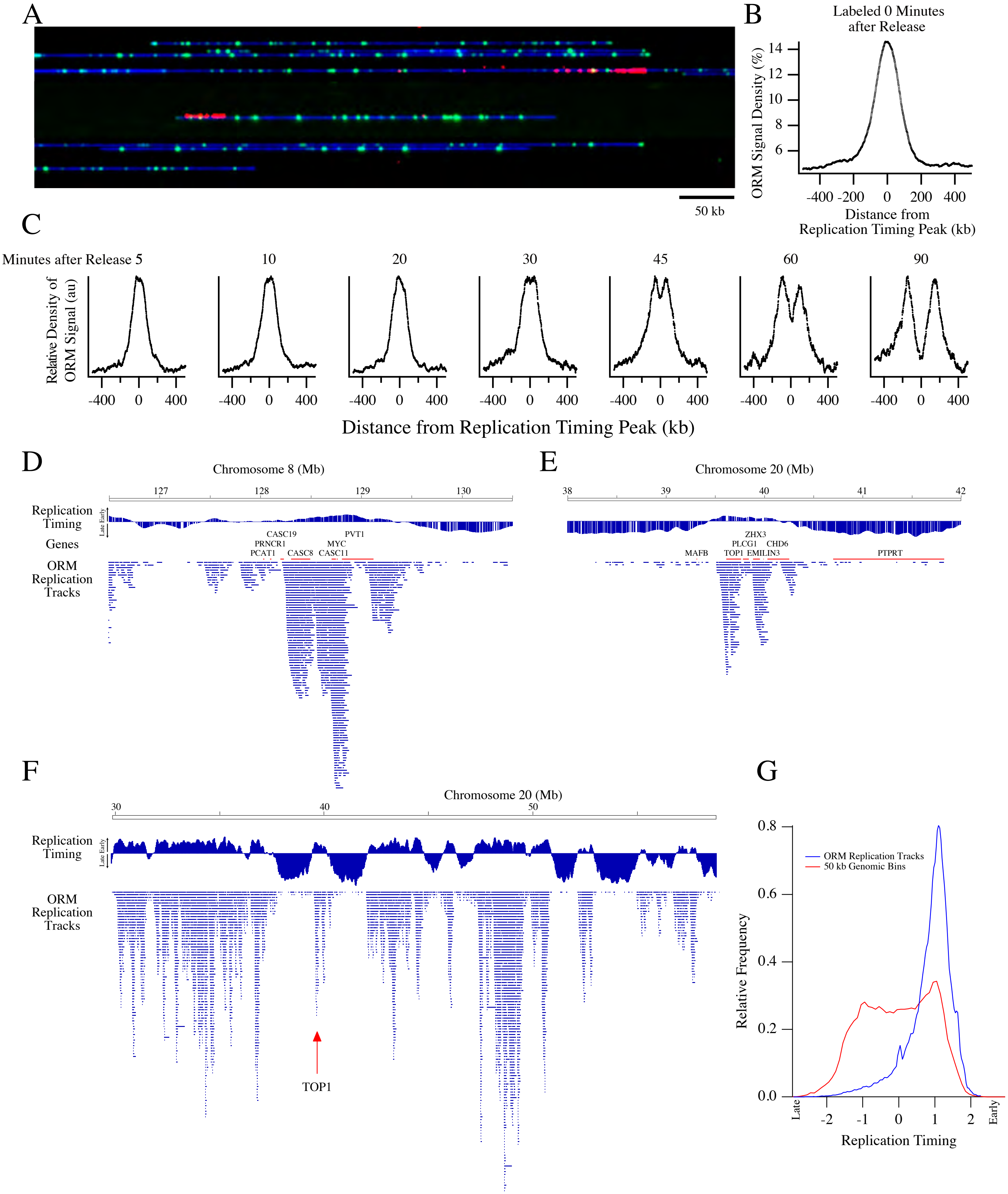
A
50 kb
B
Labeled 0 Minutes after Release
ORM Signal Density (%)
14
12
10
8
6
-400 -200 0 200 400
Distance from Replication Timing Peak (kb)
C
Minutes after Release 5
10
20
30
45
60
90
Relative Density of ORM Signal (au)
-400 0 400
Distance from Replication Timing Peak (kb)
D
Chromosome 8 (Mb)
127
128
129
130
Replication Timing
Late Early
Genes
CASC19
PVT1
PRNCR1
MYC
PCAT1
CASC8
CASC11
ORM Replication Tracks
E
Chromosome 20 (Mb)
38
39
40
41
42
ZHX3
PLCG1
CHD6
MAFB
TOP1
EMILIN3
PTPRT
F
Chromosome 20 (Mb)
30
40
50
Replication Timing
Late Early
ORM Replication Tracks
TOP1
G
ORM Replication Tracks
50 kb Genomic Bins
Relative Frequency
0.8
0.6
0.4
0.2
0.0
Late
-2
-1
0
1
2
Early
Replication Timing

Figure 1. Optical Replication Mapping of Synchronized Cells Identifies Early Initiation Events

HeLa S3 cells were synchronized by mitotic shake off, arrested in early S-phase with aphidicolin, electroporated with fluorescent dUTP and released from the aphidicolin arrest, allowing fluorescent nucleotide incorporation around early replication initiation events.

A) A representative image of the Bionano data with DNA in blue, Nt.BspQI restriction sites in green and incorporated nucleotides in red. The field of view is 750 kb.

B) Enrichment of ORM signals from the combined 0-minute dataset around early replication timing peaks, those that replicate in the first quarter of S phase.

C) Enrichment around early replication timing peaks of ORM signals from datasets in which cells were labeled between 5 and 90 minutes after release from synchronization.

D, E) Distribution of replication tracks at two previously characterized human replication initiation zones: Myc (C) and Top1(D). The distribution of replication tracks in shown beneath the HeLa S3 replication timing profile (Hansen et al., 2010) and genes from the region.

F) Distribution of replication tracks across the right arm of Chromosome 20, as in D.

G) Distribution of replication timing of replication tracks compared to replication timing of each 50 kb bin in the genome.

initiation zones at an estimated rate of 1.65±0.31 kb per minute (Figure 1C), consistent with prior estimates of mammalian replication fork rates (Jackson and Pombo, 1998; Conti et al., 2007; Chagin et al., 2016). This result is consistent with the signal in our 0' data representing initiation events and the signals after various release times representing forks moving away from such initiation events.

To facilitate visualization of labeled replication, we segmented the fluorescent-nucleotide incorporation signal into discrete replication tracks, identifying replication forks active during our labeling period (see Methods for segmentation details). In our combined 0' dataset, we observed 1,035,587 replication tracks. Because of the sparse labeling, 58% of our replication tracks are labeled with only one signal. To test whether these solo signals are sparsely labeled replication tracks or spurious background noise, we calculated the enrichment of solo signals at early replication sites. Solo signals are enriched at such sites, demonstrating that they, too, mark regions of replication initiation (Figure S2C) and that noise, if any, is very low with this method.

For visualization, we used only tracks labeled with at least two signals, which contain an average of 6.2±3.8 signals and average 38.4±25.8 kb in length. For fibers on which we identify more than one track, the average distance between track centers is 111±78 kb (Figure S2D), consistent with previous measurements (Cayrou et al., 2011; Jackson and Pombo, 1998). We find enrichment of replication tracks around previously identified sites of human replication initiation (Figure 1D,E). In particular, we examined the distribution of replication tracks around the MYC and Top1 loci, both characterized as early-firing origins in the human genome (Tao et al., 2000; Keller et al., 2002), and found a pronounced enrichment at these loci, particularly in the intergenic regions surrounding these genes. More generally, replication tracks are enriched in the earliest replicating parts of the genome, as defined by replication timing profiling (Hansen et al., 2010), as expected for early-initiation events (Figures 1F,G). Interactive display of the ORM data is available on the Nucleome Browser <http://orm.nucleome.org> (Zhu et al. in prep, personal communication, J. Ma).

We observe 1.0 million initiation events in our combined 0' dataset. From the average replication track length of 38.4 kb, we infer that 46.6 Gb of DNA is labeled, which is ~1.5% of the 3.36 Tb in the dataset. Therefore, we are labeling the DNA replicated in the first ~1.5% of S phase and during that time, about 1000 initiation events occur per genome (1,035,587 initiation events observed in 1,109 genome equivalents).

Determination of Genome-Wide Replication Kinetics in Asynchronous Cells

The experiments described above provide unprecedented single-molecule mapping of initiation events at the onset of S phase. However, most cell types are not amenable to such precise cell-cycle synchronization. To map replication kinetics in unperturbed cells, we labeled asynchronous HeLa cells and H9 human embryonic stem cells, and mapped replication incorporation tracks by ORM (Figure 2A). We identified 412,115 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 S1). Using the same analysis as for our synchronous datasets, we infer a similar nucleotide-labeling frequency of 1/1025 and 1/850 thymidines, respectively. The replication tracks average 37.1±30.0 kb in length in the HeLa data and 44.0±38.7 kb in length in the H9 data, which 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 (Figure 2B).

In order to use this data to infer replication kinetics, it is necessary to determine the polarity of each incorporation track, so that the direction of each mapped replication fork is known. To determine track polarity, we took advantage of the fact that, as the labeled nucleotides are consumed, the label density

Figure 2

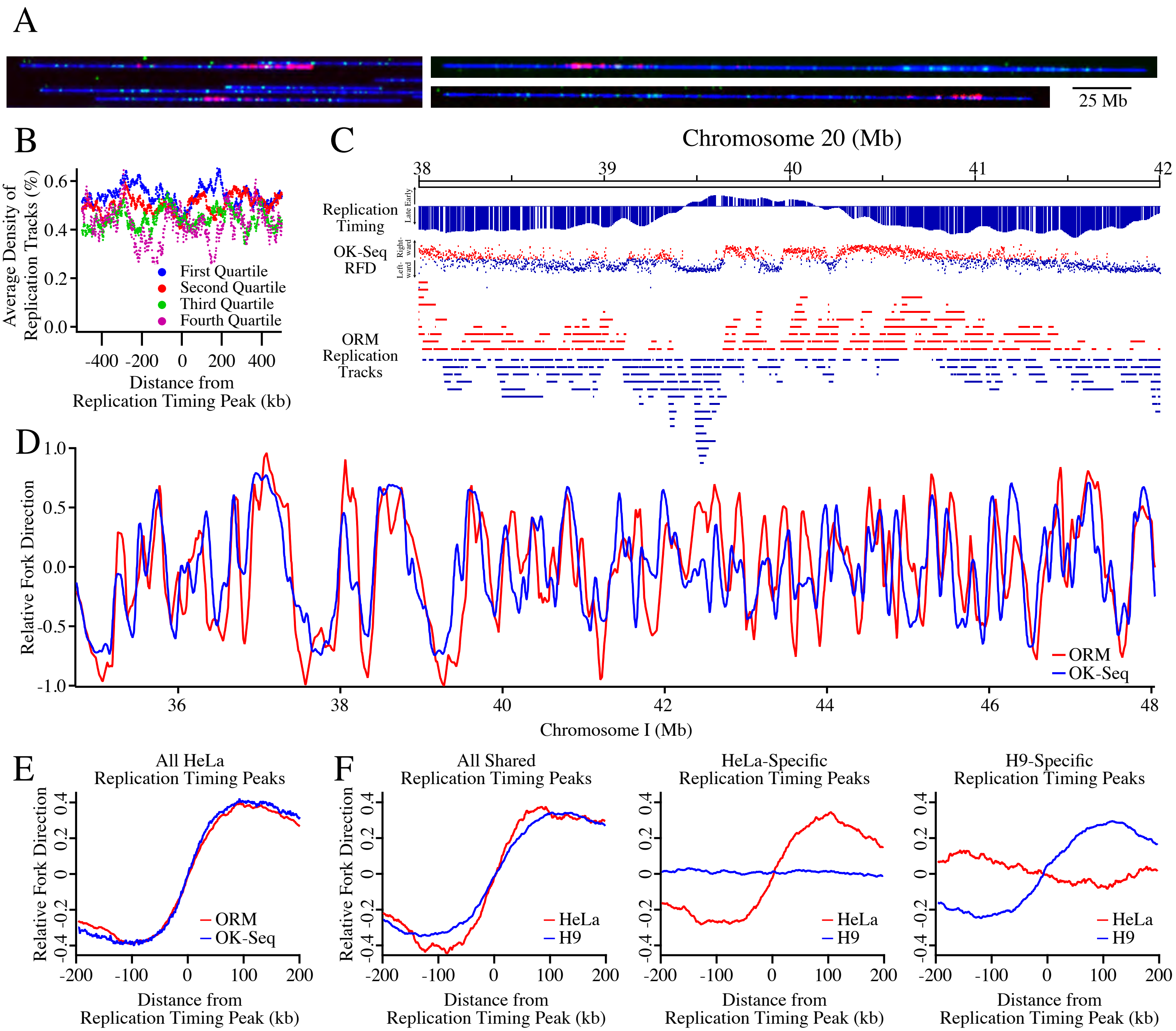
A
25 Mb
B
Average Density of Replication Tracks (%)
0.6
0.4
0.2
0.0
First Quartile
Second Quartile
Third Quartile
Fourth Quartile
-400 -200 0 200 400
Distance from Replication Timing Peak (kb)
C
Chromosome 20 (Mb)
38
39
40
41
42
Replication Timing
Early
Late
OK-Seq RFD
Right-ward
Left-ward
ORM Replication Tracks
D
Relative Fork Direction
1.0
0.5
0.0
-0.5
-1.0
ORM
OK-Seq
36
38
40
42
44
46
48
Chromosome I (Mb)
E
All HeLa Replication Timing Peaks
Relative Fork Direction
0.4
0.2
0
-0.2
-0.4
ORM
OK-Seq
-200 -100 0 100 200
Distance from Replication Timing Peak (kb)
F
All Shared Replication Timing Peaks
HeLa-Specific Replication Timing Peaks
H9-Specific Replication Timing Peaks
HeLa
H9

Figure 2. Optical Replication Mapping of Asynchronous Cells Identifies Cell-Type-Specific Genome-Wide Replication Kinetics

A) Asynchronous HeLa S3 and H9 human embryonic stem cells were electroporated with fluorescent dUTP, allowing fluorescent nucleotide incorporation into ongoing replication fork.

B) The density of ORM signals from the HeLa asynchronous dataset around replication timing peaks separated into replication timing quartiles, defined by average replication timing (S50 from Chen et al., 2010).

C) Labeled ORM replication forks, with leftward-moving forks colored blue and rightward-moving forks colored red, are depicted below the replication timing profile of the Top1 region and its OK-seq relative fork direction (RFD) data, with blue dots indicating leftward replication, red dots rightward, and the vertical position of the dot, between -1 and 1, indicating the fraction of replication moving in that direction. The polarity of ORM replication forks was determined by calculating their fork direction index (FDI).

D) The RFD calculated from ORM and OK-seq data across 7 Mb of Chromosome 1.

E) The enrichment of HeLa fork polarity, as measured by RFD, around early replication timing peaks.

F) The enrichment of HeLa and H9 fork polarity around cell-type specific and cell-type general early replication timing peaks.

diminishes, a strategy that has been used in other fiber-analysis approaches (Müller et al., 2019; Huberman and Riggs, 1968; Hennion et al., 2020). We calculated a fork direction index (FDI) as the ratio of integrated fluorescent signal in the left half of the track to that in the right half, for all tracks with 3 or more signals.

As expected, ORM replication tracks tend to be oriented in the direction of replication inferred from the slope of replication timing profiles and from Okazaki fragment mapping (Hansen et al., 2010; Petryk et al., 2016) (Figures 2C,D). Furthermore, genome-wide analysis of replication fork polarity around replication timing peaks shows that the polarity signal in ORM data is comparable to that in published OK-seq data (Figure 2E). Importantly, the polarity signal in ORM data is cell-type specific (Figure 2F). 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.

Genome-Wide Mapping of Early-Firing Human Initiation Zones

To map the genome-wide distribution of early-firing human initiation sites, we analyzed the incorporation signal in our combined 0' dataset and identified 4,628 initiation zones, defined as peaks of replication signal density (Figure 3A, see Methods for details). The initiation zones are 40±13 kb in length (Figure S2E), comparable to initiation zones mapped by OK-seq (Petryk et al., 2016). The initiations zones cover about 6% of the genome, however, they contain only about 40% of the ORM signal. Much of the rest of the ORM signal is directly adjacent to defined initiation zones, but much is distributed elsewhere at lower density, particularly in late-replicating regions on the genome (Figure 3A). Plotting the initiation efficiency of 50 kb genomic bins show that there is a continuous distribution of initiation efficiency between the higher-efficiency regions we define as initiation zones and lower-efficiency regions, with no obvious break between them (Figure S2F).

The median genome-wide inter-initiation-zone distance is 294 kb. However, since our dataset is biased towards early replicating initiation zones, we suspect that we underestimate the number of initiation zones in late-replicating parts of the genome. Consistent with that contention, limiting our analysis to the earliest quarter of the genome leads to a median inter-initiation-zone distance of 174 kb. The inter-initiation-zone distances in the other three quarters of the genome are 270, 362 and 319 kb, respectively. We suspect that the lower density of observed initiation zones in the later-replicating regions of the genome results from having less data rather than fewer initiation events in these regions (Figure 1G).

A major advantage of ORM is that it records the number of fibers that both have and do not have replication tracks, allowing for the calculation of the absolute efficiency with which replication initiates at each locus in the genome. Although such an analysis is possible with prior methods, it is considerably more difficult and thus rarely reported (Demczuk et al., 2012). Such an analysis shows that early initiation is quite heterogeneous. Examining the 10% of initiation zones that have the earliest average replication times, we find that they initiate replication, on average, on only 16% of molecules (Figure 3B). We therefore conclude that, in aphidicolin arrested cells, replication initiates in only a stochastically-selected subset of initiation zones, including only16% of the most efficient early initiation zones.

We investigated the frequency of early initiation of replication across the genome, examining all 4,628 ORM initiation zones. We find that the local efficiency of initiation in these zones is generally below 20%, with occasional efficiencies up to 40% and a median efficiency of 6.5% (Figure 3B). 97% of the initiation zones with efficiencies of more than 10%, and 95% of the zones with efficiencies of more than 5%, are in early-replicating regions of the genome (Figure 3B), consistent with the results in Figure 1G. To directly test whether the initiation frequency we measure by ORM is correlated with the initiation efficiency measured by OK-seq (Petryk et al., 2016), we compared the two initiation frequencies within each OK-seq initiation zone. We find a significant correlation between the two

Figure 3

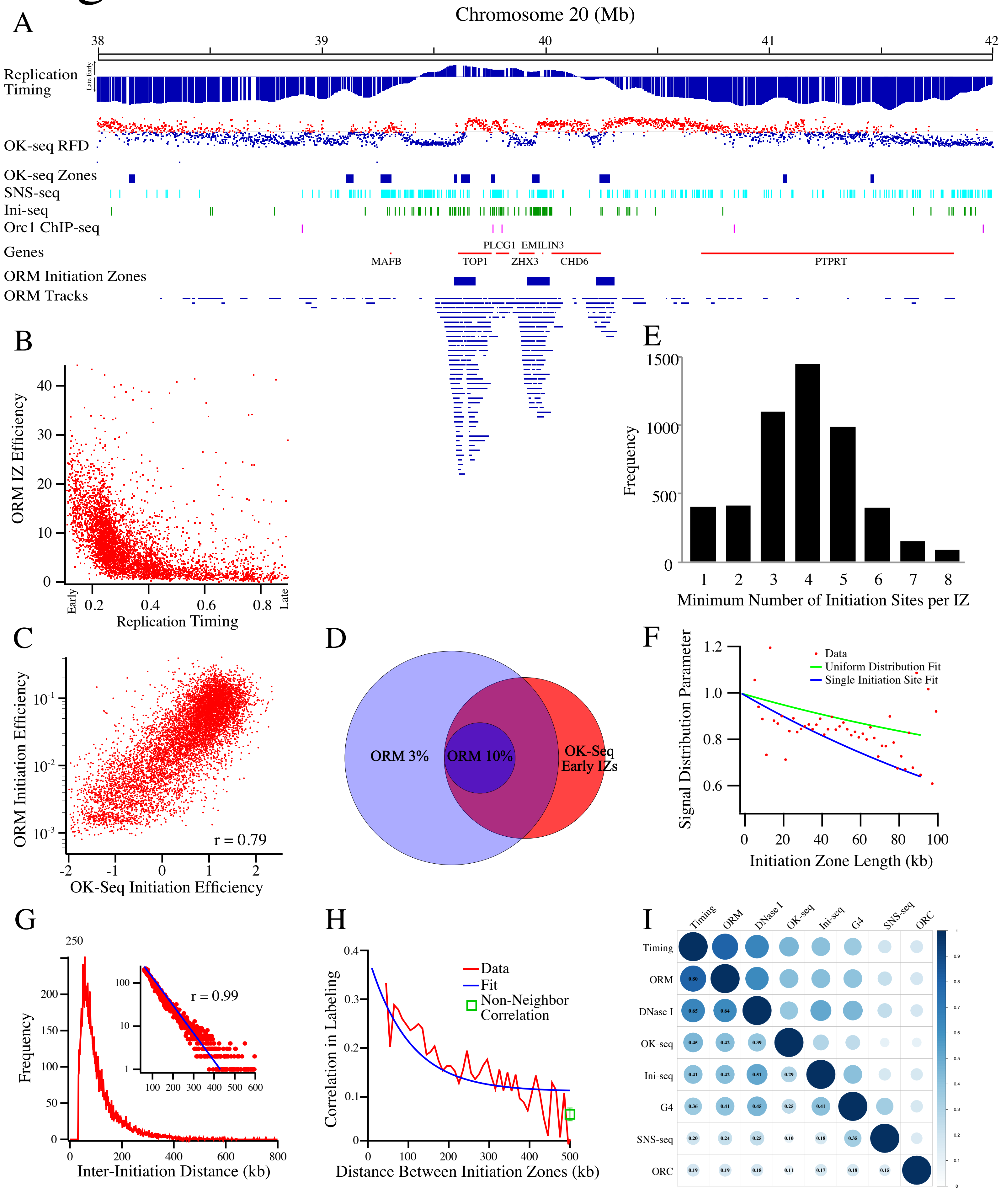
A
Chromosome 20 (Mb)
38
39
40
41
42
Replication Timing
Early
Late
OK-seq RFD
OK-seq Zones
SNS-seq
Ini-seq
Orc1 ChIP-seq
Genes
PLCG1
EMILIN3
MAFB
TOP1
ZHX3
CHD6
PTPRT
ORM Initiation Zones
ORM Tracks
B
ORM IZ Efficiency
40
30
20
10
0
Early
0.2
0.4
0.6
0.8
Late
Replication Timing
E
1500
1000
500
0
Frequency
1
2
3
4
5
6
7
8
Minimum Number of Initiation Sites per IZ
C
ORM Initiation Efficiency
10^{-1}
10^{-2}
10^{-3}
r = 0.79
-2
-1
0
1
2
OK-Seq Initiation Efficiency
D
ORM 3%
ORM 10%
OK-Seq Early IZs
F
Signal Distribution Parameter
1.2
1.0
0.8
0.6
Data
Uniform Distribution Fit
Single Initiation Site Fit
0
20
40
60
80
100
Initiation Zone Length (kb)
G
250
200
150
100
50
0
Frequency
100
10
1
r = 0.99
100
200
300
400
500
600
0
200
400
600
800
Inter-Initiation Distance (kb)
H
0.4
0.3
0.2
0.1
0
Correlation in Labeling
Data
Fit
Non-Neighbor Correlation
0
100
200
300
400
500
Distance Between Initiation Zones (kb)
I
Timing
ORM
DNase I
OK-seq
Ini-seq
G4
SNS-seq
ORC
0.80
0.65
0.64
0.45
0.42
0.39
0.41
0.42
0.51
0.29
0.36
0.41
0.45
0.25
0.41
0.20
0.24
0.25
0.10
0.18
0.35
0.19
0.19
0.18
0.11
0.17
0.18
0.15
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0

Figure 3. Mapping of Human Replication Initiation Zones

A) A comparison of ORM tracks ORM initiation zones, OK-seq RFD (as in Figure 2C), OK-seq initiation zones (Petryk et al., 2016), SNS-seq (Picard et al., 2014), Ini-seq (Langley et al., 2016) and Orc1 ChIP-seq (Dellino et al., 2013) at the Top1 locus.

B) The correlation between ORM initiation efficiency (fraction of labeled fibers/IZ) in the combined 0-minute dataset and replication timing (S50 from Chen et al., 2010) in the 4,628 ORM initiation zones.

C) The correlation between ORM initiation efficiency and initiation efficiency inferred from OK-seq (Petryk et al., 2016) data in the 9,832 OK-seq initiation zones.

D) A Venn diagram of the overlap between OK-seq early initiation zones and 50 kb windows in the genome that have ORM initiation efficiencies of at least 3% or 10%.

E) The estimated minimal number of initiation sites per initiation zone. The estimate was made by calculating minimum number of replication tracks in each initiation zone whose centers are more than 15 kb apart. Most of the IZs for which all track centers are with in 15 kb of each other are small and contain few replication tracks (Figure S2B).

F) The distribution of signal across initiation zones. 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. 21) than if initiation is distributed across the IZ (Eq. 24, Supplemental Mathematical Methods). See Supplemental Mathematical Methods for details.

G) 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.

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

I) A matrix depicting the correlation between initiation sites predicted by ORM, replication timing (Chen et al., 2010), OK-seq (Petryk et al., 2016), SNS-seq (Picard et al., 2014), Ini-seq (Langley et al., 2016) Orc1 ChIP-seq (Dellino et al., 2013), DNase I hypersensitivity (Bernstein et al., 2012) and G4 motifs (Puig Lombardi et al., 2019).

(Figure 3C, r = 0.79), consistent with early initiation zones being defined by higher initiation efficiency in early S phase. In a parallel analysis, we find that 68% of the 6,166 early initiation zones identified by OK-seq show at least 3% initiation efficiency by ORM and 100% of loci that show greater than 10% efficiency by ORM are identified by OK-seq as initiation zones (Figure 3D).

Visual inspection of the distribution of replication tracks suggests that initiation in human cells is not concentrated at discrete, frequently-used origins. Instead, initiation appears to be distributed across broad initiation zones, as has been observed at specific loci and in other genome-wide analyses (Petryk et al., 2016; Mesner et al., 2013; Besnard et al., 2012; Hamlin et al., 2008; Dijkwel et al., 2002; Demczuk et al., 2012; Lubelsky et al., 2011; Tao et al., 2000). Even the initiation zone at the Top1 locus appears to have initiation events distributed over tens of kilobases (Figures 1E, S3A). However, given the sparse labeling of our replication tracks, we cannot identify initiation sites with resolution higher than about 15 kb. Nonetheless, 93% of initiation zones contain incorporation tracks with centers that are more than 15 kb apart, suggesting that these initiation zones contain at least two, and possibly many more, initiation sites (Figures 3E and S3). Furthermore, the 7% of initiation zones that do not have obviously non-overlapping replication initiation sites tend to be smaller (26.8 kb v. 40.1 kb), less efficient (3.7% v. 7.9%) and later replicating (S50 0.58 v. 0.35) than general (Figure S3B).

Another way to distinguish between initiations that occur at a single site and those that are distributed across a broad region is to examine the distribution of ORM signals across the initiation zones, the expectation being that a single initiation site will have narrower distribution than distributed initiation sites. We measured the signal at the edges of initiation zones as a fraction of the signal at the middle of the initiation zone and fit to that data two models, one based on a single initiation site and another based on a uniform distribution of initiation sites across the initiation zone (Figure 3F). Neither model fits well, with most of the data falling between the model predictions. From this result, we conclude that neither extreme accurately describes the distribution of initiation sites and that, most likely, initiation events are distributed at several to many discrete sites throughout initiation zones. Taking all of our data together, we find no evidence of isolated or high-efficiency initiation sites in the human genome.

Initiation Events in Neighboring Initiation Zones are not Clustered

Regardless of how initiation events are locally distributed, it is important to know whether local initiation events are coordinated. A popular hypothesis in the field is that initiation events are clustered on chromosomes, such that initiation events will occur simultaneously in neighboring initiation zones (Fragkos et al., 2015). However, this hypothesis has never been rigorously tested, because the single-molecule experiments that have identified multiple initiation events on individual chromosomes have never quantitated their frequency relative to single initiation events in same initiation zone (Jackson and Pombo, 1998; Cayrou et al., 2011; Lebofsky et al., 2006; Huberman and Tsai, 1973; Blow et al., 2001; Marheineke and Hyrien, 2004), which is necessary to determine if multiple initiation events happen more often than would be expected by chance.

With our deep, single-molecule data on long (up to megabase-length) fibers, we can quantitatively test whether initiation events in neighboring initiation zones are correlated. A simple, unbiased test for correlation between events is to determine whether the spacing between the event deviates from an exponential distribution. Items distributed randomly on a line have an exponential distribution of inter-item distances (Birnbaum, 1954); therefore, correlated clusters of initiation events would deviate from an exponential distribution of inter-initiation distances. We measured 18,275 inter-initiation distances on 13,769 fibers containing multiple initiation events (Figure 3G). We measure few initiation events less than 60 kb apart, due to the facts that the tracks themselves average 40 kb in length and that early initiation zones have a median distance of 174 kb. However, above 60 kb, there is an exponential distribution of inter-initiation distances (r = 0.99, Figure 3G, inset). This distribution indicates that, on this length scale, initiation events are distributed randomly.

As an orthogonal test of whether initiation events in neighboring initiation zones are positively or negatively correlated, we compared the frequency of labeling at each initiation zone with the frequency with which its neighboring initiation zones were labeled on the same fiber (Figure 3H). Neighboring initiation zones can be co-labeled as a result of individual initiation in each zone or of initiation in one zone leading to passive incorporation of label in a neighboring zone by the same replication fork. Correlation can also arise because of heterogeneous uptake of labeled nucleotides. Basically, if one cell has a very high level of nucleotide uptake, all initiation zones (and any two neighboring initiation zones, in particular) will be more likely to incorporate label, making them appear to be correlated (see Methods for a more detailed explanation). Therefore, we fit a model to our initiation-zone co-labeling data that takes into account the previously determined incorporation track length (Figure S2B) and nucleotide uptake heterogeneity (Figure S1B), and assumes that individual initiation events are independent. This model fits the data well, suggesting there is no measurable positive or negative correlation between initiation in neighboring zones (Figure 3H). Furthermore, any apparent correlation due to heterogeneity of nucleotide uptake would be seen among all initiation zones in a cell, not just neighboring ones. Thus, we also calculated the correlation coefficient among all initiation zones on each fiber, not just the neighboring ones, and found a coefficient similar to that of neighboring initiation zones separated by more than 200 kb, consistent with the interpretation that the low level of residual correlation seen at longer distances is a global effect, not a local spatial correlation (Figure 3H).

Comparison of Different Origin Mapping Datasets

We compared our optical replication mapping 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). By visual inspection, replication tracks appear to be enriched around the initiation zones identified by OK-seq (Figure 3A). Across the genome, we see colocalization of OK-seq initiation zones and ORM signal. We see similar colocalization of ORM signal with Ini-seq signal, a cell-free initiation mapping approach, Orc1 ChIP-seq peaks, and, to a lesser extent, SNS-seq peaks, which maps initiation events by sequencing short nascent strands produced by replication initiation (Figure 3A). 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 data sets (Figure 3I). We find that ORM replication tracks correlate to a similar extent with OK-seq and Ini-seq (r = 0.42) and to a lesser extent with SNS-seq (r = 0.24) and Orc1 ChIP-seq (r = 0.19).

Characterization of Initiation Zones Mapped by ORM

To characterize the genomic character and chromatin context of our initiation zones, we investigated the enrichment of various genomic and epigenomic annotations (Figure 4). Comparing our IZs with various chromatin states from the ENCODE project (Bernstein et al., 2012; Ernst and Kellis, 2012), we find that they are enriched in enhancers and low activity intergenic regions and depleted in transcription units and polycomb-repressed chromatin (Figure S4A). Examining specific histone modifications, our IZs are enriched in H2AZ, which has been implicated in ORC binding, and enhancer-specific modifications, such as H3K4me1 and H3K27ac, and they are depleted in, although adjacent to, transcription-elongation-associated modifications, such as H3K79me2 and H3K36me2 (Figures 4 and S4B,D), in agreement with previous studies (Petryk et al., 2016; Long et al., 2020; Pourkarimi et al., 2016). Consistent with the enrichment of initiation events in enhancer regions and their depletion in transcribed regions, IZs are enriched in DNase I hypersensitive sites and depleted in RNA pol II (Figures 4 and S4B,D).

Figure 4

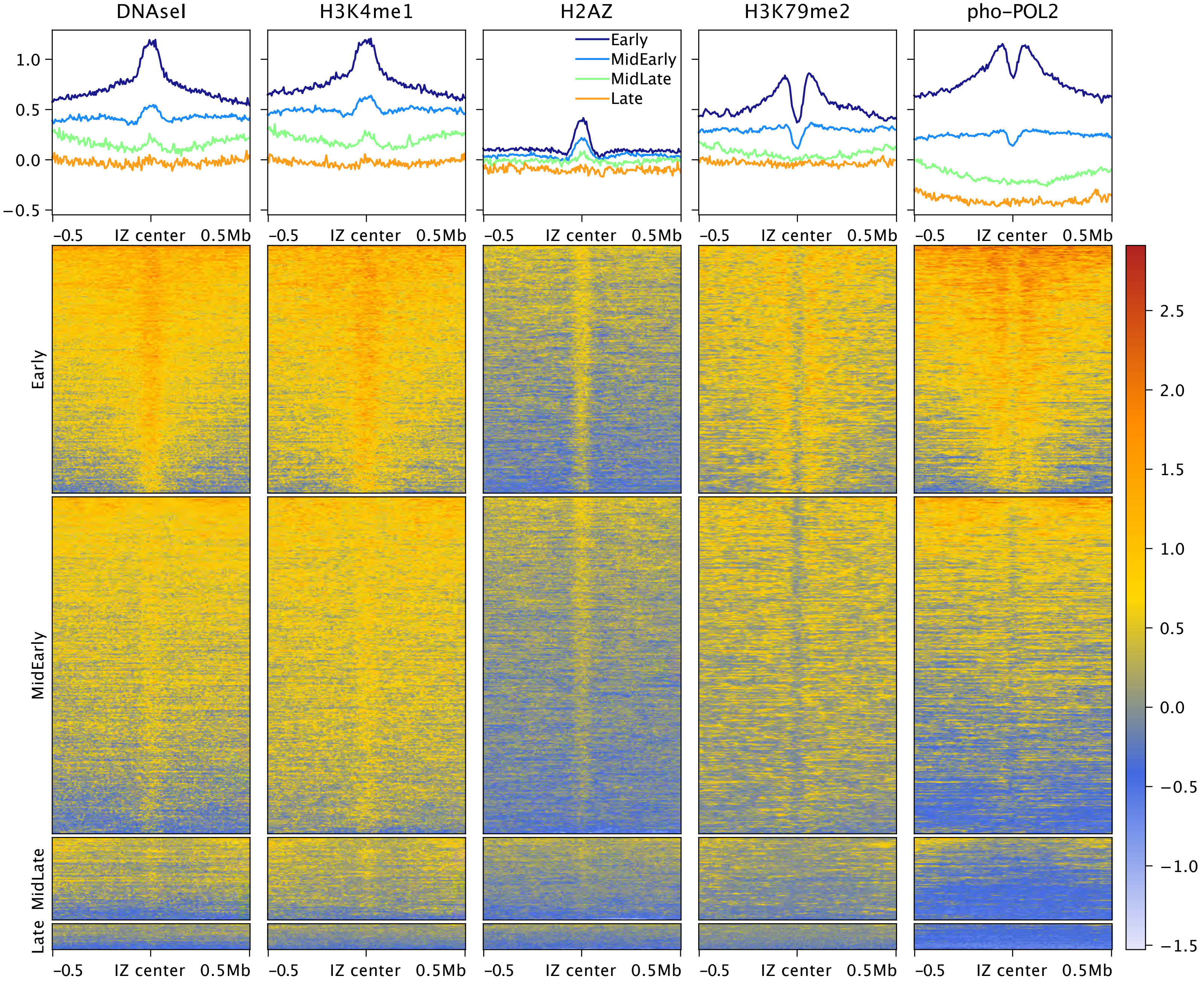

Figure 4. Characterization of Human Initiation Zones

Enrichment of histone modifications and other genomic features relative to ORM initiation zones, separated into replication timing quartiles. The upper panels show the genome-normalized relative signal around all IZs. The lower panels show heat maps of the same signal at each IZ. The y-axis scale for the upper panels and the heat-map scale for the lower panels are the same genome-normalized enrichment in relative units.

Although metazoan ORC has only a weak affinity for AT-rich sequences *in vitro* (Vashee et al., 2003; De Carli et al., 2018) and no discernible sequence preference *in vivo* (Miotto et al., 2016), sequences that form G4 quadruplex structures have been proposed to correlate with initiation sites in vertebrate cells (Cayrou et al., 2015; Langley et al., 2016; Valton et al., 2014; Prorok et al., 2019). We tested whether the initiation zones we identified are enriched in G4 motifs, using a GC-content-adjusted background model to avoid detecting G4 enrichment as a trivial consequence of the GC-rich nature of enhancers. Of the 737,736 G4 sequences computationally identified in the human genome (Puig Lombardi et al., 2019), 78,796 are found in initiation zones, 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 initiation zones) outside of initiation zone are no less likely to contain G4 motifs (176,077 with G4 and 71,614 without) than GC-rich regions within initiation zones (10,371 with G4 and 7,314 without; chi-square test, p = 0.078). Therefore, although we find that IZs are GC rich (Figure S4C), 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, per se.

Early Initiation Events in Late-Replicating Domains

Although 91% of our early initiation events are in early-replicating regions of the genome, 9% occur in late-replicating domains (Figures 1F, 3B). By visual inspection, the replication tracks in late-replicating parts of the genome appear to be in the locally earliest-replicating parts of the late-replicating regions (for instance at 56 and 58 Mb in Figure 1F), suggesting that they reflect the location of the initiation zones normally used to replicate these regions and not contaminating asynchronous cells. To test this suggestion, we calculated the enrichment of replication tracks around initiation zones identified as replication timing peaks (Chen et al., 2011) grouped into four quartiles, defined by average replication timing (S50 from Chen et al., 2010). We find significant enrichment of replication tracks initiation zones in first three quartiles, (but no enrichment in the fourth quartile, which contains only 3% of the genome-wide replication timing peaks), demonstrating that our ORM signal in late-replicating regions is associated with initiation zones (Figure 5A). Moreover, the signal-to-noise ratios in the first three quartiles is similar (SNR = 3.2-2.8, Figure 5A), demonstrating that, although there is much less signal in the third quartile, it is enriched over background to the same extent as in the earlier quartiles. In contrast, no enrichment is seen in any quartile in asynchronous cells (Figure 2A). These results confirm that the enrichment of early initiation events we see at late-replicating initiation zones cannot be attributed to contaminating asynchronous background signals and, therefore, presumably results from bona fide early initiation events in initiation zones that, on average, replicate late in S phase.

Computational Modeling Suggests that Early- and Late-Replicating Initiation Zones are Distinguished Only by Relative Initiation Probability

The distribution of early initiation events we observe—frequent firing in early-replicating initiation zones and rare firing at late-replicating initiation zones—is reminiscent of the stochastic regulation of origin-firing timing observed in budding yeast (Czajkowsky et al., 2008; Yang et al., 2010; de Moura et al., 2010). Models of stochastic replication timing posit that the only difference between early- and late-firing origins is their relative probability of firing, and that this relative probability does not change during S phase. More-efficient origins are more likely to fire first and thus, on average, have earlier firing times, but sometimes they still fire late, whereas less-efficient origins are less likely to fire early and thus, on average, have later firing times, but do sometimes fire early (Rhind et al., 2010).

Such models predict that, because the relative probability of origin firing does not change during S phase, measuring the relative probability of initiation in early-S phase is sufficient to determine the kinetics of initiation for all initiation zones throughout S phase. If this prediction holds true, we should be able to calculate the complete replication timing profiles of cells even though we have assayed firing

Figure 5

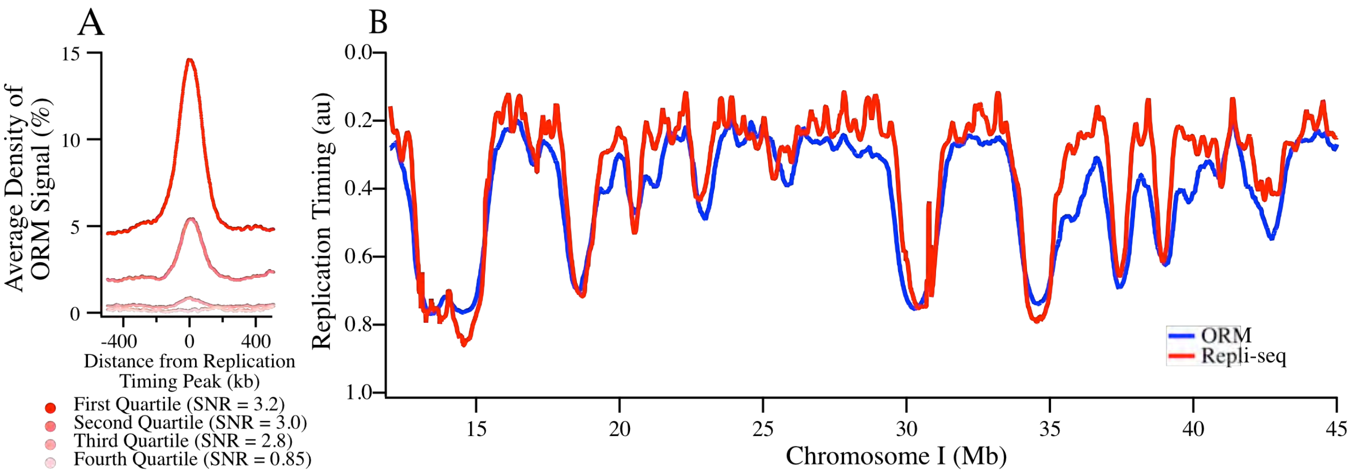

Figure 5. Probability of Early Initiation Predicts Replication Kinetics Genome-Wide
A) Enrichment of ORM signals from the combined 0-minute dataset around replication timing peaks separated into replication timing quartiles. Signal-to-noise ratio (SNR) was calculated as maximum signal over average signal between 400 and 500 kb away.
B) Comparison between experimentally determined replication timing (S50 from Chen et al., 2010) and replication timing predicted from ORM data using a stochastic model (Gindin et al., 2014a), r = 0.81.

probabilities in only the first 1.5% of S phase. To test this prediction, we used Replicon, a stochastic replication simulator that calculates replication profiles from a so-called Initiation Probability Landscape (IPLS), the probability of initiation at each point in the genome (Gindin et al., 2014a). We find that we can, indeed, accurately recapitulate genome-wide replication timing profiles with our ORM data (r = 0.81, Figure 5B). Further, our ORM data predicts replication timing significantly better than other initiation data sets, such as, OK-Seq (r = 0.73), Ini-seq (r = 0.71) or SNS-seq (r = 0.52, Figure S5). ORM also predicts replication kinetics slightly better than DNase I hypersensitivity (r = 0.80), which had previously been identified as the best predictor of initiation (Gindin et al., 2014a). These result support the hypothesis that the primary difference between early- and late-replicating initiation zones is their relative initiation probability, influenced by chromatin accessibility, and not a qualitative difference in their initiation times.

Discussion

We have developed Optical Replication Mapping (ORM), which, for the first time, allows genome-wide, high-throughput, single-molecule replication mapping in human cells. The approach is based on *in vivo* labeling of replication forks and *in vitro* detection and genomic mapping of incorporated label on long, single DNA fibers using the Bionano Saphyr platform. Single flow cells using the current mapping approach yielded an average 150-fold coverage of the human genome on fibers from 150 kb to 2 Mb, with an average length of 300 kb (Table S1). Altogether, we have collected 2,550-fold coverage with an average fiber length of 284 kb. Our method is similar to the recently described HOMARD approach, which was used to map DNA replication in *Xenopus* egg extracts (De Carli et al., 2018).

Direct, single molecule, measurements of initiation frequency

Detection of DNA synthesis on individual DNA fibers is the most direct way to localize sites of DNA replication. Since the preponderance of the literature describes low-frequency, heterogeneous selection of replication initiation sites in mammalian cells, single-molecule measurements may be the only means to accurately map the relative frequencies of initiation events in the human genome. However, to date, single-molecule replication-mapping technologies, such as DNA combing and SMARD, which have throughput of a few hundred fibers, have provided insufficient depth of coverage for metazoan genome-wide analyses. Recently, nanopore sequencing has been used to map replication incorporation tracks genome-wide on single DNA fibers, but the current technology only allows for relatively small datasets of relatively short fibers (Georgieva et al., 2019; Müller et al., 2019; Hennion et al., 2020). The largest dataset from a recent study describing the DNAscent approach to nanopore-based replication mapping in budding yeast was 3.8 Gb (the equivalent of just over 1x coverage of the human genome) with an average fiber length of 32 kb (Müller et al., 2019). Therefore, ORM, for which current typical datasets are 150-times larger, provides a unique opportunity for single-molecule, genome-wide mapping of metazoan replication at high coverage on long fibers.

Using ORM, we have mapped the replication initiation events occurring in the first 1.5% of S phase in aphidicolin-synchronized HeLa cells. These initiation events, as others have reported (Petryk et al., 2016; Long et al., 2020; Pourkarimi et al., 2016; Ganier et al., 2019; Cayrou et al., 2015), are enriched in the enhancer regions of active genes and are enriched in active chromatin modifications, in particular DNase I HS, H2AZ and H3K4me1 (Figure 4). However, unlike previous reports, we can directly estimate the frequency with which initiation occurs in each initiation zone because we measure all fibers, both labeled and unlabeled. We find that the distribution of early initiation events is heterogeneous. Although 99% of early initiation zones predicted by OK-seq analysis are detected in our ORM data, only 6% of them, on average, initiate at the beginning of S phase in any given cell. Furthermore, although most of the initiation events that we observe are in early-replicating initiation zones, a small but significant minority are in late-replicating initiation zones, and the frequency of initiation is correlated with the replication timing of the initiation zone (Figure 3B). This observation is not due to contamination of our synchronized G1/S cells with non-synchronized cells because the signal in asynchronous cells shows no enrichment at specific sites, let alone previously mapped late initiation sites (Figure 2A). These results suggest that, in any given cell, a small number of potential initiation sites are stochastically selected from a large number of potential initiation sites, and that the probability of a site being selected is correlated with the replication time of that site, with early-replicating sites having a higher probability of initiating than late-replicating sites (Figure 3B).

We have mapped replication initiation sites across the genome with over 1,000-fold genome coverage, allowing us to identify initiation sites used in as few as 0.1% of cells. The initiation events we map appear to be spread over broad initiation zones, consistent with previous observations (Petryk et al., 2016; Mesner et al., 2013; Besnard et al., 2012; Hamlin et al., 2008; Dijkwel et al., 2002; Demczuk et al., 2012; Lubelsky et al., 2011; Tao et al., 2000; Anglana et al., 2003). However, the sparse labeling and relatively long length (~40 kb) of our incorporation tracks allow us estimate the location of initiation sites at a resolution of only about 15 kb. Nonetheless, in 93% of initiation zones, we observe incorporation track centers that are more than 15 kb apart, suggesting that they initiated from distinct sites (Figure 3E). Furthermore, the distribution of signal across initiation zones is not generally consistent with a single initiation site per zone (Figure 3F). We cannot distinguish between frequent, discrete, origins scattered throughout each initiation zone and diffuse initiation zones in which initiation can occur almost anywhere. However, we can rule out the hypothesis that human replication initiates at isolated, efficient, well-defined replication origins. In particular, we see no evidence for unique initiation sites at the lamin B or Top1 loci, which have been proposed to be isolated origins (Keller et al., 2002; Abdurashidova et al., 2000). We suspect that the inefficient nature of human initiation zones and the low signal-to-noise ratio of previous initiation-mapping technology made it difficult to comprehensively map initiation sites across initiation zones. However, our data suggests that human initiation zones contain multiple low-efficiency initiation sites, as has been previously suggested (Langley et al., 2016; Petryk et al., 2016; Mesner et al., 2003; Hamlin et al., 2008).

Human Replication Initiation Events are not Clustered

Our single-molecule data also allows us to directly measure correlations between initiation sites on individual chromosomes. A number of previous reports have claimed that initiations events are clustered, in that initiation events are more likely to occur next to one another on chromosomes than would be expected by chance, and that these clustered events are regularly spaced between 75 and 150 kb apart in human cells (Jackson and Pombo, 1998; Cayrou et al., 2011; Lebofsky et al., 2006; Huberman and Tsai, 1973; Blow et al., 2001; Marheineke and Hyrien, 2004). However, these observations have been largely anecdotal because they have recorded observations only on fibers that have multiple initiation events, and so have not rigorously measured their frequency relative to single initiation events. To test whether initiation events tend to occur in clusters, we examined the frequency and distribution of multiple tracks on single fibers (Figure 3G). If events on a line are clustered with a characteristic inter-event distance, the distribution of inter-event distances will be bell-shaped, with a peak at the most frequent inter-event distance. If, however, the events are distributed randomly, with no inter-event correlation, the distribution of inter-event distances will be exponential (Birnbaum, 1954). We find that, between 60 and 600 kb, the distribution of inter-initiation-event distances is exponential (r = 0.99), inconsistent with any significant clustering of initiation.

As an orthogonal test of initiation clustering, we directly examined co-replication of neighboring initiation zones. We examined every pair of initiation zones on each fiber we collected and asked whether the frequency with which they were replicated was correlated (Figure 3H). Using a model that takes into account both passive replication predicted by incorporation track lengths (Figures S2B) and nucleotide-uptake heterogeneity (Figure S1B) but otherwise assumes no correlation among initiation events, we can account for the observed frequency of replication in neighboring initiation zones (Figure 3H), consistent with the conclusion that human replication initiation sites are not clustered in early S phase. Thus, although initiation events with similar replication timing tend to be clustered because regions of similar replication timing are grouped together (Figure 1F), initiation events are no more likely to occur next to one another, or in collinear clusters, than expected by chance (Figures 3G,H).

Stochastic Usage of Initiation Zones in Human Cells

The heterogeneous, uncorrelated nature of replication initiation that we observe is reminiscent of stochastic models that have been proposed to explain the regulation of replication timing in budding yeast. Such models propose that the average timing of initiation in a population is regulated by the probability of a site initiating in an individual cell. Sites with a high initiation probability are, on average, more likely to initiate early and thus have an earlier average replication time, although sometimes they initiate late. Sites that have a lower initiation probability are, on average, less likely to initiate early and thus have a later average replication time, although sometimes they initiate early. In particular, neither hierarchical regulation of initiation timing nor correlation between initiation at neighboring sites is necessary in such models. Previous analyses of human replication kinetics have suggested positive correlations, in which one initiation event, or the forks it produces, would increase the probability of other nearby initiation events (Guilbaud et al., 2011; Löb et al., 2016). However, a more recent, higher resolution replication timing study found no evidence of this so-called "domino model" (Zhao et al., 2020). In any case, we find that a stochastic model assuming independent origin initiation can explain our observed data (Figure 3H).

To test whether our results are compatible with a rigorously stochastic model of replication kinetics, we used a computational model of DNA replication that assumes stochastic origin firing (Gindin et al., 2014a). Using our initiation-site mapping data to initialize the model produces predicted replication timing profiles that give an excellent fit (r = 0.81) to experimentally produced profiles (Figure 5B).

Since the computational model does not contain any correlation between initiation events, this analysis suggests that any positive or negative correlations between initiation events must be of sufficiently low magnitude or restricted spatial distribution as to be not required to explain observed genome-wide patterns of replication timing. Our work, in combination with work from budding and fission yeast (Kaykov and Nurse, 2015; Yang et al., 2010; Patel et al., 2006; de Moura et al., 2010), supports the hypothesis that stochastic control of initiation timing is a fundamental feature of DNA replication, conserved across eukaryotes.

General Application of ORM to Map Replication Kinetics

In addition to mapping replication initiation sites, we also used ORM to map ongoing replication forks in unperturbed, asynchronous cells. We inferred fork direction from the asymmetric nature of incorporation tracks caused by the reduction of incorporation as labeled nucleotides are depleted, an approach that has been used before for both radio- and fluorescently labeled nucleotides (Müller et al., 2019; Huberman and Riggs, 1968; Hennion et al., 2020). Averaged over our nearly-500-fold fiber coverage of the genome, our inferred fork directions allow us to infer genome-wide replication kinetics that agree with independently determined replication-timing and replication-fork-directionality profiles (Figures 2C,D). In future work, it should be possible to use sequential labeling with nucleotides of different colors to unambiguously determine fork direction, thereby greatly increasing the resolution and sensitivity of asynchronous ORM. In addition to identifying replication initiation sites throughout the S phase of any cell type, sequential labeling will allow investigation of other important aspects of replication. 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. Such application will make ORM a central technique for studying DNA replication, DNA repair and genome instability.

Methods

Cell Labeling and Sample Preparation

HeLa S3 cells were grown in DMEM plus 10% Cosmic Calf Serum (GE Life Science SH30087) and Pen/Strep. For asynchronous HeLa datasets, cells were grown to ~70% confluency, washed with warm PBS, 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). For G1/S synchronized data sets, cells were synchronized at metaphase by incubation with 0.05 μg/ml nocodazole for 4 hours followed by shake off. The percentage metaphase cells was quantified by metaphase spread. Briefly, 3X volumes of ddH20 was added to an aliquot of media from shake off and was incubated at 37˚C for 15 minutes. Cells were fixed by adding several drops of ice-cold methanol:glacial acetic acid (3:1). Cells were spun down and pellet were resuspended in ice-cold methanol:glacial acetic acid (3:1) and spotted on to a clean slide. Slides were allowed to dry, spotted with Vectasheild (Vector Laboratories H-1000) containing DAPI and covered with coverslips. Metaphase and interphase cells were counted on a Nikon Eclipse Ti microscope. Metaphase cells were released into fresh media containing 10 μg/ml aphidicolin (Millipore Sigma 178273) for 16 hours. For 0-minute release datasets, cells were washed three times with warm PBS (or PBS plus aphidicolin, in the case of B.0) to remove aphidicolin, 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). Cells were recovered in warm media plus aphidicolin for 30 minutes in a 37˚C water bath with agitation every 5 minutes to prevent cells from settling in the tube. Cells were then spun down and washed 3 times with warm PBS to remove aphidicolin. For timed release experiments, cells were washed three times with warm PBS to remove aphidicolin, warm media was added, and cells were returned to 37˚C/5% CO2 incubator for the indicated times (5, 10, 20, 30, 45, 60, or 90 minutes). Cells were then washed again with PBS, trypsinized, and electroporated in the presence of fluorescent nucleotides as above. In all cases, cells were allowed to recovered overnight in media at 37˚C/5% CO2. After recovery, all cells were detached with trypsin and snap frozen in liquid nitrogen.

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. Cells were grown to ~70% confluency, washed with warm PBS, dissociated to single cells with Gentle Cell Dissociation Reagent (StemCell Technologies 07174) and electroporated (Lonza, Nucleofection kit P3, hESC program) in the presence of 40 µM Aminoallyl-dUTP-ATTO-647N (Jena Bioscience NU-803-647N). Cells were recovered overnight in StemPro media plus 10uM Y-27632 dihydrochloride (ROCK inhibitor, Sigma Y0503) at 37˚C/5% CO2.

Nucleotide labeling efficiency was assayed by plating a small aliquot of cells on coverslips after electroporation. Following overnight recovery, cells were fixed with 4% formaldehyde for 10 minutes at room temperature. Cells were then washed twice with PBS followed by permeabilization and counterstaining with PBS plus 0.2% Triton X-100 plus DAPI for 5 minutes at room temperature. Coverslips were washed twice with PBS, spotted with Vectasheild, and placed on slides. Images were captured on a DeltaVision (GE Life Sciences) microscope. Replication foci patterns were categorized as previously described (O'Keefe et al., 1992).

Optical Replication Mapping

Two Bionano mapping approaches were used: Nick, Label, Repair and Stains (NLRS) and Direct Label and Stain (DLS) (Table S1). For NLRS, cell pellets were thawed on ice, resuspended in cold Cell Buffer (Bionano 80004) and embedded into low-melting point agarose plugs using the CHEF Mammalian Genomic DNA Plug Kit (Bio-Rad #1703591) according to the manufacturer's instructions, with a target concentration of 1,000,000 cells per plug. Cells immobilized in plugs were lysed and treated with Proteinase K and RNase A. The plugs were then melted and solubilized, and the resulting DNA was further cleaned by drop dialysis. The DNA was fluorescently labeled with green fluorophores at Nt.BspQI sites and counterstained with YOYO-1 using Bionano Genomics' NLRS labeling kit (Bionano 80001). Throughout the entire process, the samples were shielded from light whenever possible, in order to maintain integrity of the fluorophores. For DLS, DNA was labeled with green fluorophores using the direct labeling enzyme DLE-1 (Bionano Genomics 80005) according to manufacture's protocol. The rest of processing steps were the same as NLRS labeling.

Labeled DNA samples were loaded onto a Saphyr chip (Bionano Genomics 20319) and run on the corresponding Saphyr instrument (Bionano Genomics IN-011-01), following the manufacturer's instructions. Briefly, the DNA was coaxed into an array of parallel nanochannels via electrophoresis, thereby elongating the DNA to a uniform contour length, allowing for accurate measurement of distances along the molecules. Molecules were imaged to collect the YOYO-1 DNA signal in a false-color blue channel, the Nt.BspQI or DLE-1 labels in the green channel, and the *in vivo* replication incorporation tracks in the red channel.

Images were converted to digitized molecules and the positions of the green and red labels on each fiber were determined using Bionano Access. The Nt.BspQI and DLE-1 labels were aligned to human reference genome hg19 by creating an *in silico* digestion of the reference FASTA file. Alignment of the fibers' Nt.BspQI and DLE-1 sites allowed mapping of the red replications signal to the genome. The pipeline produces one .bnx file containing the intensity and location of each green and red signal on each fiber and one .xmap file containing the genomic location of each fiber. All primary data and analysis scripts are available by request.

Modeling Signal-Intensity and Inter-Signal-Distance Distributions

To model the signal-intensity distributions (Figure S2A), we assumed that signals represent individual fluorophores that can be resolved with some finite resolution. We modeled both the photon output and the inter-signal distance (which controls the probability of having two fluors closer together than the resolution of the Saphyr) of each fluor as Poisson processes, resulting in the depicted fit (obtained from the first four terms of Eq. 5, Supplemental Mathematical Methods), which predicts that 20% of the fluors are closer than the resolution of the Saphyr. See Supplemental Mathematical Methods for details.

To model the distribution of inter-signal distances (Figure S2B), we assume that the incorporation rate is a Poisson process for which the Poisson statistic is determined by the average amount of label taken up by cells. We further assume that there is an exponentially decreasing amount of label over time as it is consumed by replication. Combining these two assumptions results in the depicted fit (Eq. 14, Supplemental Mathematical Methods), with an inferred initial incorporation rate of about 1/900 thymidines and a depletion half length of about 75 kb for the merged 0-minute dataset. See Supplemental Mathematical Methods for details.

Combining the models in Figures S2A and B, we calculate that 20% of the signals in Figure S2B (corresponding to the 20% of the fluors that we infer from Figure S2A are not individually resolved by the Saphyr) are located within 1.3±0.3 kb of a neighboring fluor, consistent with the reported resolution of the Saphyr and the observation that there are almost no inter-signal distances below 1.3 kb in our datasets.

Identification of Replication Tracks

After mapping the fibers on the human genome, the distances between the genomic mapping signals were recalibrated to match their reference genome positions, compensating for heterogeneity in fiber stretching. We then mapped signals in the red channel between two adjacent genomic mapping signals as nucleotide-incorporation signals. The nucleotide-incorporation signal is discontinuous and appears as clusters of neighboring signals (Figure 1A). To segment the ORM signals into replication-incorporation tracks, we used a two-step process based on Gaussian mixture models. We first fit a Gaussian mixture model using the R package *mixtools* <https://www.rdocumentation.org/packages/mixtools> with 3-gaussian distributions to the inter-signal distance distribution, with the first peak corresponding to the distances between signals within the same segment, the second peak corresponding to the distances between signals in adjacent segments and the third peak corresponding to signals in non-adjacent segments. The tail of the first Gaussian distribution (16,384 kb) was comparable amongst various data sets and was therefore selected as the cutoff value (cutoff1) to merge adjacent signals into primary segments. We then measured the distance between adjacent primary segments and fit a 2-Gaussian distribution to find the second cutoff value (cutoff2, 32,768 kb), which we used to merge all the adjacent primary segments into final replication tracks.

To estimate the resolution with which we can identify the initiation site of any given replication track, we defined the maximum likelihood distribution of initiation sites for a given replication track and, using numeric simulations, estimated the standard deviation of that distribution as 14.2±0.1 kb. See Supplemental Mathematical Methods for details.

To determine the direction of replication tracks in our asynchronous data, we defined a fork direction index (FDI): FDI = log2(IntensitySUM_Left / IntensitySUM_Right), where IntensitySUM_Left and IntensitySUM_Right is the total fluorescent signal in the left and right half, respectively, of each ORM track with at least 3 signals. A positive (negative) FDI corresponds to a rightward (leftward) moving replication fork. The mean relative fork direction (RFD) within a given window is defined as (R-L)/(R+L), where R and L is the number of rightward (FDI>0) and leftward (FDI<0) ORM tracks, respectively.

Identification of Initiation Zones

Initiation zones were identified based on the ORM signal density along the genome. We first obtained the ORM signal density by calculating the percentage of fibers containing ORM signal in 10 kb sliding windows with 1 kb steps. LOESS smoothing (with $\alpha = 0.75$ and the polynomial degree = 2) was then performed in successive 160 kb windows, with 8 kb overlaps to avoid discontinuity of fitting. Within the 8 kb overlaps, the final smoothed values were the averages, weighted by their distance to the corresponding window, of the smoothed values from two adjacent windows. Initiation zones were then defined as the peaks on the smoothed ORM signal density profile. The size of each initiation zones was defined by the smallest window that containing at least 40% of ORM signals within each peak. The value of 40% was chosen because, when aggregated around IZ peaks, the ORM signal density decreases with distance from the peak at a similar rate until about 40% then decreases much more slowly. This process was performed on all 4 biological replicates of HeLa 0' data as well as the combined dataset. Only initiation zones identified in all 4 replicates and with a relative peak height in the combined dataset (as measured by the difference in signal density between the peak and the edges of the zone) of greater than 0.015%, or those identified in at least 2 replicates and with a relative peak height of greater than 0.05%, were retained.

To estimate the minimum number of initiation sites in each initiation zone, we calculated number of replication track centers separated by more than the resolution of our ability to localize initiation sites, which is about 15 kb. We calculated the position of the center of each replication track that overlaps a given IZ, discarding any that are more than 7.5 kb outside of it. We then ranked the centers by position and, starting at one end, counted the number of centers that are more than 15 kb apart.

Comparative Analysis of ORM Data with Other Datasets

The raw Repli-Seq data of HeLa S3 cells were downloaded from the Encode project <http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeUwRepliSeq> (Hansen et al., 2010) and S50 (the fraction of S phase at which 50% of the DNA is replicated in a defined genome region) was computed and used to identify replication timing peaks as previous described (Chen et al., 2011). Published H9 ESCs replication timing and peak calling data were used (Zhao et al., 2020). ORM signal enrichment around timing peaks was calculated in R (version 3.5.1 <https://www.r-project.org>).

Correlations between ORM initiation efficiency and published genome-wide HeLa cell 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)—were calculated across the genome in 50 kb adjacent windows. When necessary, genomic coordinates were remapped to hg19 using LiftOver <https://genome.ucsc.edu/cgi-bin/hgLiftOver>. ORM initiation efficiency was calculated as the number of fibers with ORM signal in a 50 kb genomic window divided by the number of fibers covering that window. The OK-seq initiation efficiency was calculated either genome-wide using the previously described origin efficiency metric (OEM) approach (McGuffee et al., 2013) or for each OK-Seq initiation zone with the ΔRFD between its right and left extremities (Petryk et al., 2016). Correlations were calculated using Spearman's correlation coefficient in R.

The genome-segmentation results (ChromHMM) based on ENCODE data were retrieved from UCSC genome browser <https://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeAwgSegmentation>. To compensate for the large size of IZ, relative to many of the ChromHMM segments, we used the IZ's core regions, defined as the region from center of the IZ to the point of highest ORM signal within that IZ.

Histone modification data for HeLa S3—H2A.Z, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27ac, H3K27me3, H3K36me3, H3K79me2 and H4K20me1—was downloaded from ENCODE <https://www.encodeproject.org>. If available, the data obtained from G1 phase cells and the average of all replicates were use in our analysis. The data of signal enrichment for RNA Pol II-pS2 <https://www.encodeproject.org/experiments/ENCSR000ECT> and DNase I-seq data (UW DNase I HS data track) for HeLa S3 were also downloaded from ENCODE. The signal enrichment data of each mark were re-normalized such that the background regions show zero log2 enrichment ratio. Published G4 sequence locations were retrieved (Puig Lombardi et al., 2019). We analyzed 3 G4 datasets (G4L3, G4L7 and G4L12) with similar results; we therefore used the G4L12 ($G_3N_{1-12}G_3N_{1-12}G_3N_{1-12}G_3$), which contains the largest dataset and give the best correlation with ORM data, in our final analysis. The genome-wide pairwise correlation analysis and enrichment around ORM IZs were performed by deepTools (Ramírez et al., 2016) or with R (version 3.5.1 <https://www.r-project.org>). Initiation zones were classified by timing—early (S50<0.25), mid-early (0.25 ~ 0.5), mid-late (0.5~0.75) and late (S50>0.75)—using the S50 of the IZ center. The partial correlation analysis between ORM initiation efficiency and G4 sequences with controlling the effect of GC percentage (obtained from UCSC <http://hgdownload.soe.ucsc.edu/goldenPath/hg19/gc5Base>) was performed by using the R package ppcor <https://cran.r-project.org/web/packages/ppcor/index.html>. Features were visualized in IGV (Robinson et al., 2011) and the Nucleome Browser <http://orm.nucleome.org> (Zhu et al. in prep, personal communication, J. Ma).

Correlation of Initiation in Neighboring Initiation Zones

To calculate the degree of correlation between initiation events in neighboring initiation zones, we defined an initiation event as any ORM signal in any initiation zone on a fiber. We then calculated the correlation as the frequency with which we observed initiation in neighboring IZs on the same fiber divided by the product of the frequency of overall initiation in the two IZs. The non-neighbor correlation was calculated in the same manner, but using all pairwise combinations of IZs on each fiber farther that 200 kb apart. To calculate the expected correlation due to passive replication of one IZ due to initiation in a neighboring IZ, we used the observed label-depletion rate and inferred the frequency with which initiation at one IZ would lead to incorporation of label at a neighboring IZ at a given distance away (Figure S2B). To calculate the correlation expected from heterogeneity of label uptake by cells, we used the observed distribution of cellular labeling (Figure S1B). The two calculations were combined to produce the model (Eq. 30, Supplemental Mathematical Methods) shown in Figure 3H. See Supplemental Mathematical Methods for details.

Replication Timing Simulation

To simulate the replication timing profiles, we used the Replicon simulation code (Gindin et al., 2014a). Replicon uses three sets of used-defined parameters. The first set of parameters is the initiation probability landscape (IPLS), the relative probability of initiating at any point in the genome. In our simulations, we used the ORM the signal distribution or other genomic features with a bin size of 0.5 kb. The second set of parameters control the virtual flow sorter. In agreement with the results from the original paper (Gindin et al., 2014b), we set this to {0.0, 0.17, 0.35, 0.58, 0.92, 0.99, 1.0}, although other choices give similar results. The final parameter is the number of forks. Here, we again followed the original paper and set the number of forks equal to 10.24 + x(7.9E-7), where x is the length of the chromosome in base pairs.

Acknowledgements

KP, JB and NR were funded by NIH grant GM125872. CLC was supported by the I. Curie YPI program, the ATIP-Avenir program from CNRS, Plan Cancer from INSERM, the CNRS 80|Prime interdisciplinary program, ANR and INCa. WW was supported by a COFUND IC-3i International PhD fellowship.

Author Contributions

Conceptualization: CLC, DG, NR, **Methodology:** WW, KK, KP, AH, JB, CLC, DG, NR, **Investigation:** WW, KK, KP, HY, **Software:** WW, KP, MC, XZ, CLC **Validation:** WW, KK, KP, AH, HY, JB, CLC, DG, NR, **Formal Analysis:** WW, KP, JB, CLC, NR, **Resources:** AH, CLC, DG, NR, **Data Curation:** WW, CLC, **Writing – Original Draft:** NR, **Writing – Review & Editing:** WW, KK, KP, AH, HY, JB, CLC, DG, NR, **Visualization:** WW, KP, CLC, NR, **Supervision:** JB, CLC, DG, NR, **Project Administration:** NR, **Funding Acquisition:** AH, CLC, DG, NR

Figure S1

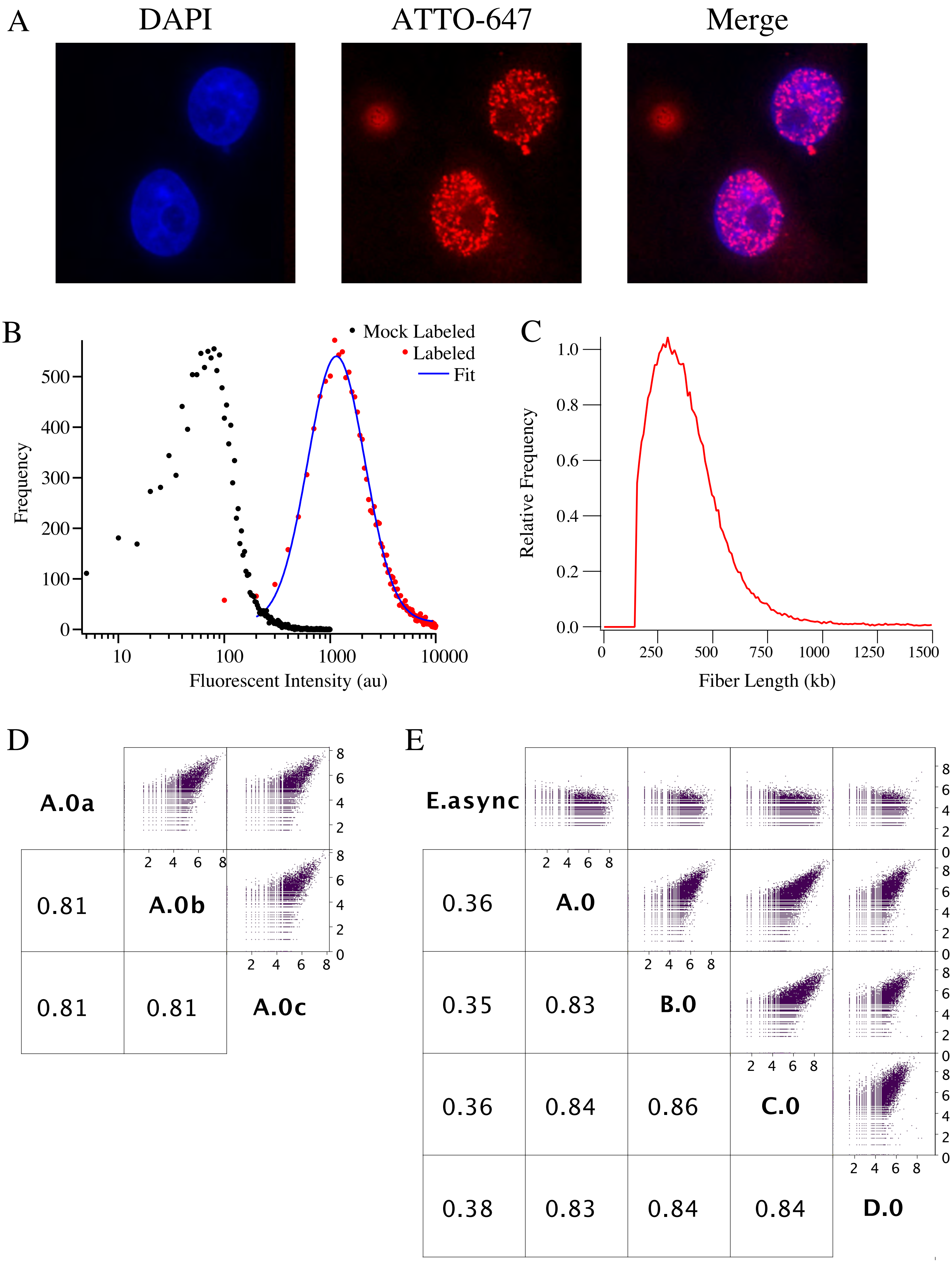
A
DAPI
ATTO-647
Merge
B
Mock Labeled
Labeled
Fit
Frequency
0
100
200
300
400
500
10
100
1000
10000
Fluorescent Intensity (au)
C
Relative Frequency
0.0
0.2
0.4
0.6
0.8
1.0
0
250
500
750
1000
1250
1500
Fiber Length (kb)
D
A.0a
A.0b
A.0c
0.81
0.81
0.81
2 4 6 8
0 2 4 6 8
E
E.async
A.0
B.0
C.0
D.0
0.36
0.35
0.36
0.38
0.83
0.84
0.83
0.86
0.84
0.84
2 4 6 8
0 2 4 6 8

Figure S1. Characterization of ORM Labeling

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.

B) 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.

C) The distribution of fiber lengths of 0-minute dataset B.0, which was collected with the DLS mapping approach and has slightly longer fibers that datasets collected with the NLRS approach (Table S1). The average fiber length is 338 kb and the fiber N50 is 369 kb.

D) 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.

E) 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, as in panel D. 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.

Figure S2

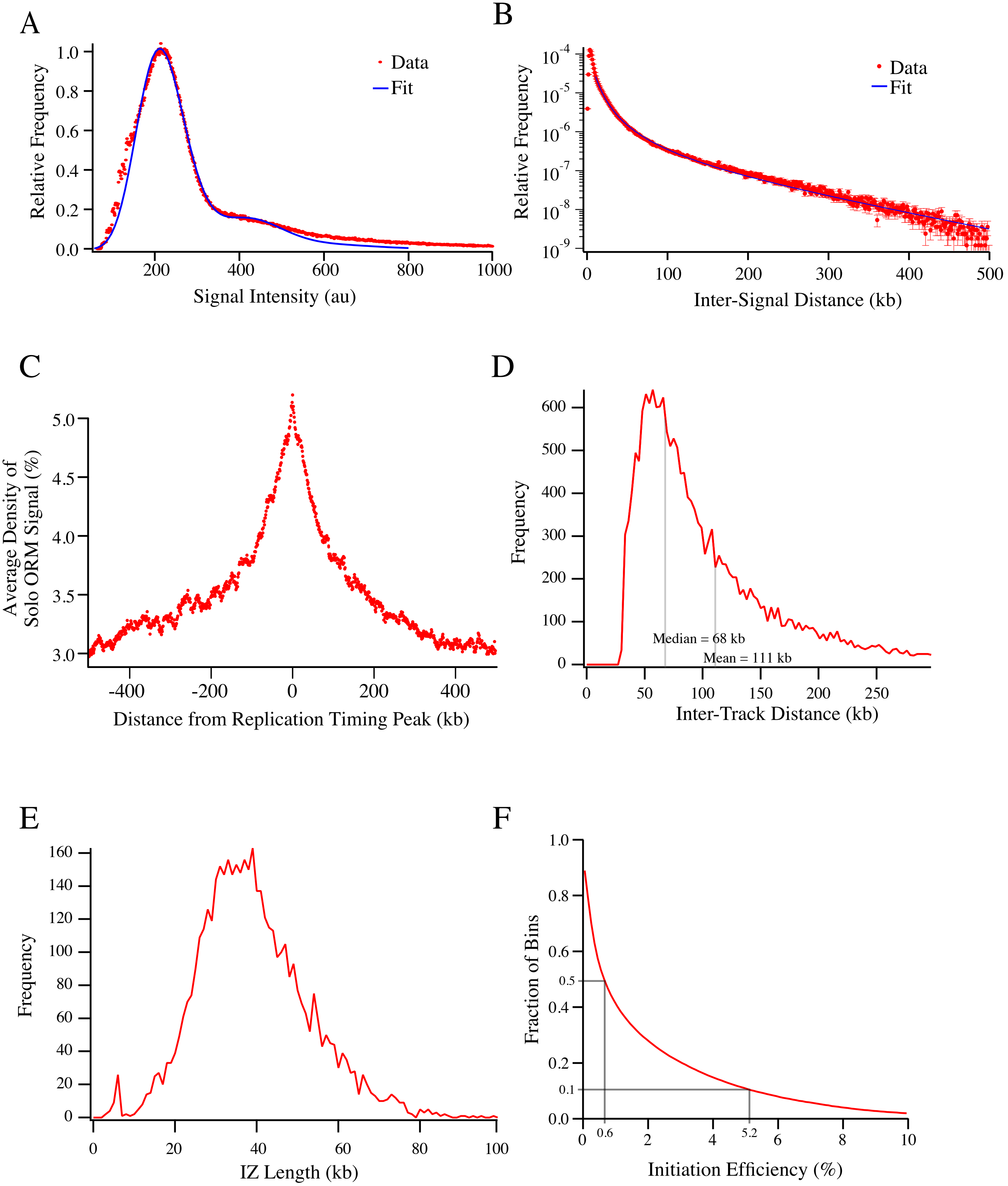
A
1.0
0.8
0.6
0.4
0.2
0.0
Relative Frequency
Data
Fit
200
400
600
800
1000
Signal Intensity (au)
B
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
10^{-9}
Relative Frequency
Data
Fit
0
100
200
300
400
500
Inter-Signal Distance (kb)
C
5.0
4.5
4.0
3.5
3.0
Average Density of Solo ORM Signal (%)
-400
-200
0
200
400
Distance from Replication Timing Peak (kb)
D
600
500
400
300
200
100
0
Frequency
Median = 68 kb
Mean = 111 kb
0
50
100
150
200
250
Inter-Track Distance (kb)
E
160
140
120
100
80
60
40
20
0
Frequency
0
20
40
60
80
100
IZ Length (kb)
F
1.0
0.8
0.6
0.5
0.4
0.2
0.1
0.0
Fraction of Bins
0
0.6
2
4
5.2
6
8
10
Initiation Efficiency (%)

Figure S2. Characterization of ORM Label Distribution

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, 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 (see panel B and Methods). 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.

B) The distribution of inter-signal distances in the combined 0-minute dataset. The fit to the data (Eq. 14, 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).

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.

D) The distribution of inter-replication-track lengths.

E) The distribution of initiation zone sizes.

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 0.6% and 10% of bins have an initiation efficiency of greater that 5.2%.

Figure S3

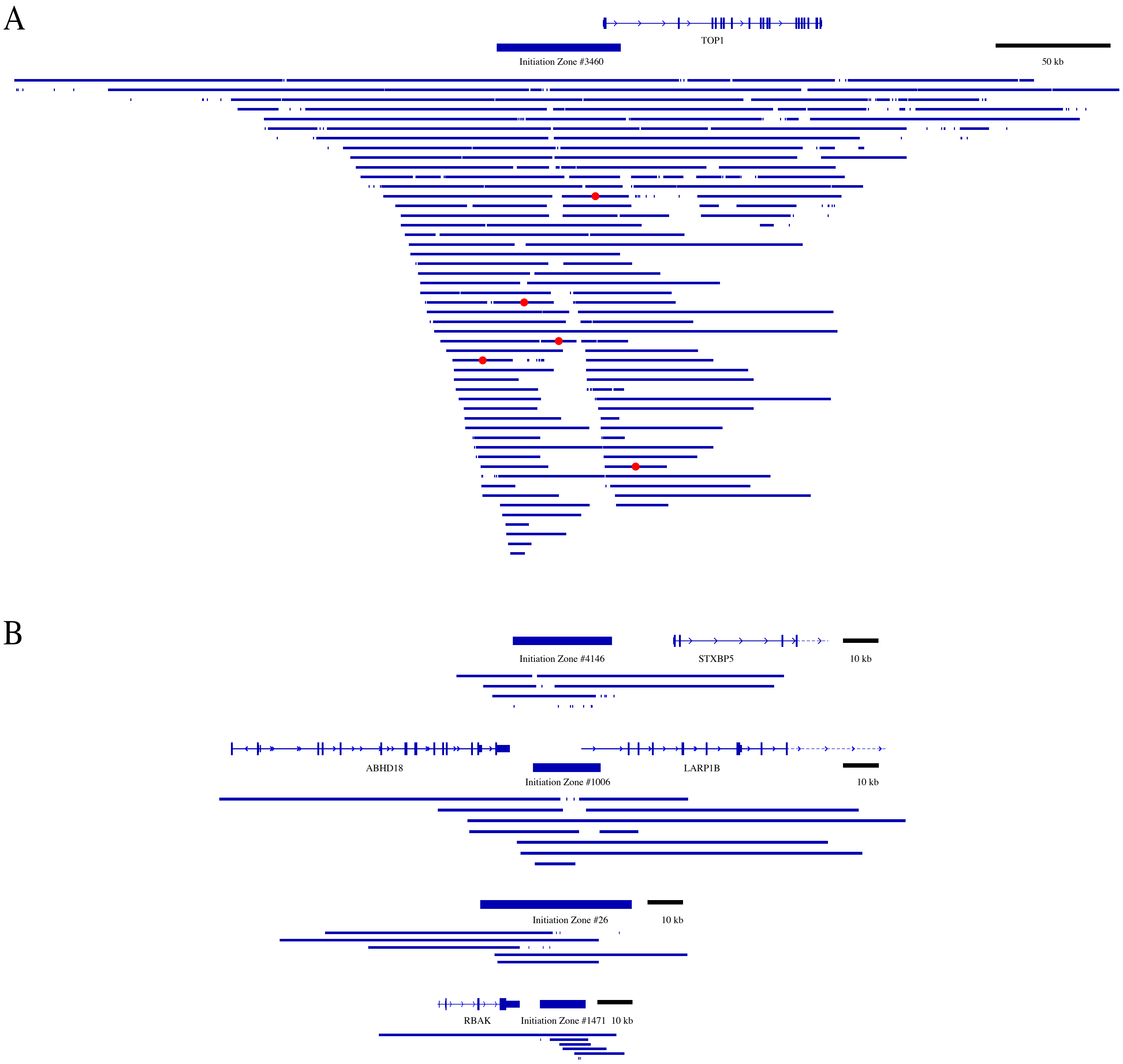

Figure S3. Distribution of Replication Tracks within Initiation Zones

A) The distribution of replication tracks in the merged 0 minute dataset at the Top1 locus. The Top1 initiation zone 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.

Figure S4

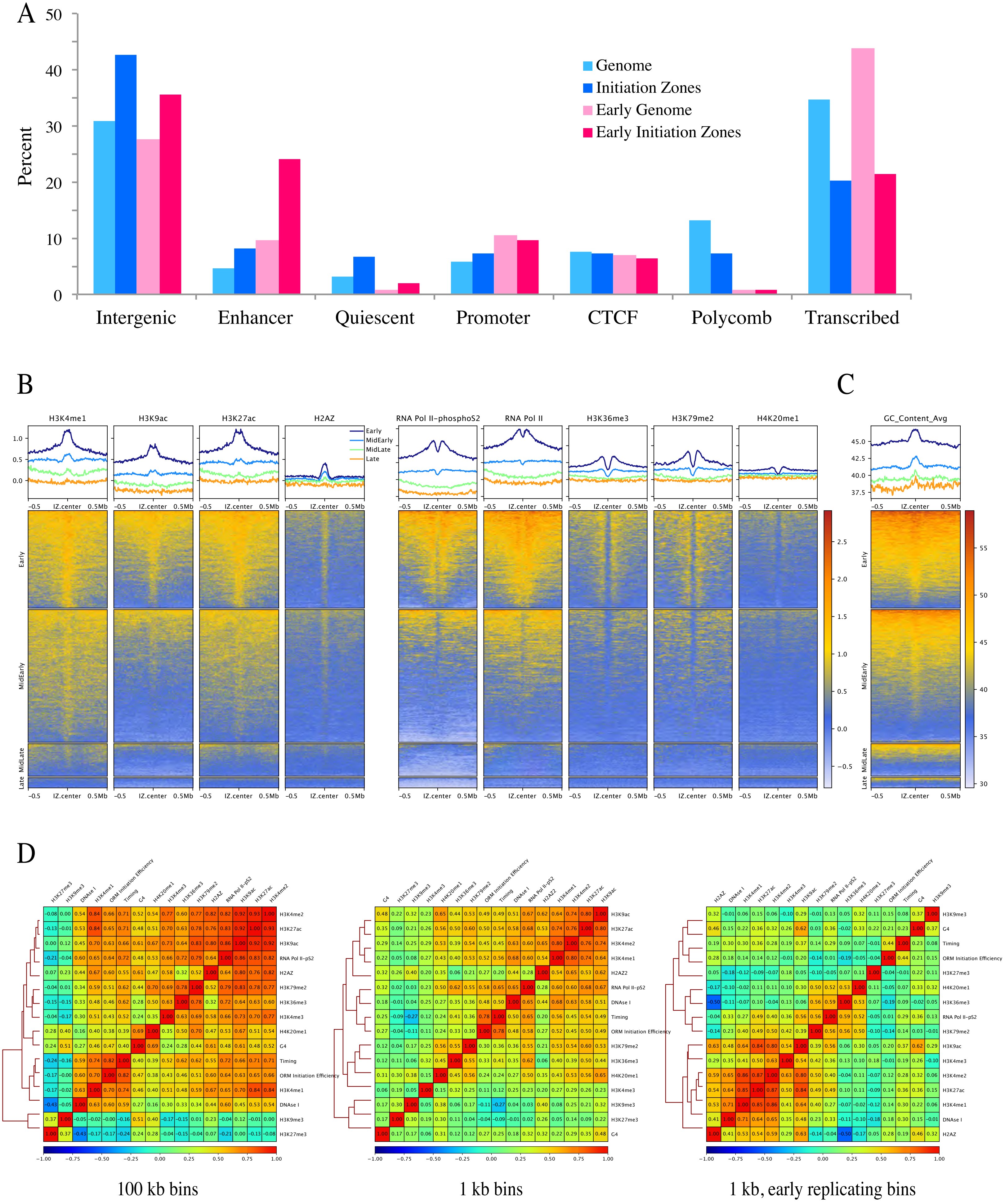
A
Percent
50
40
30
20
10
0
Genome
Initiation Zones
Early Genome
Early Initiation Zones
Intergenic
Enhancer
Quiescent
Promoter
CTCF
Polycomb
Transcribed
B
H3K4me1
H3K9ac
H3K27ac
H2AZ
RNA Pol II–phosphoS2
RNA Pol II
H3K36me3
H3K79me2
H4K20me1
Early
MidEarly
MidLate
Late
−0.5
IZ.center
0.5Mb
C
GC_Content_Avg
D
H3K4me2
H3K27ac
H3K9ac
RNA Pol II–pS2
H2AZ
H3K79me2
H3K36me3
H3K4me3
H4K20me1
G4
Timing
ORM Initiation Efficiency
H3K4me1
DNAse I
H3K9me3
H3K27me3
100 kb bins
1 kb bins
1 kb, early replicating bins

Figure S4. Enrichment of Histone Modifications and Other Genomic Features in Initiation Zones

A) The enrichment of chromatin regions defined by ChromHMM in IZs relative to the genome in general (Ernst and Kellis, 2012). The faction of IZs or genomic sequence with each of the indicated annotation is shown of for all IZs and IZs that replicate in the first quarter of S phase (S50 < 0.25).
B) The enrichment of histone modifications and other genomic features relative to ORM initiation zones. The upper panels show the genome-normalized relative ChIP-seq signal around all IZs. The lower panels show heat maps of the same signal at each IZ. The left panel includes enhancer-enrich features; the right panel shows transcription-enriched features.
C) The enrichment of GC content relative to ORM initiation zones. 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.
D) 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 correlate 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.

Figure S5

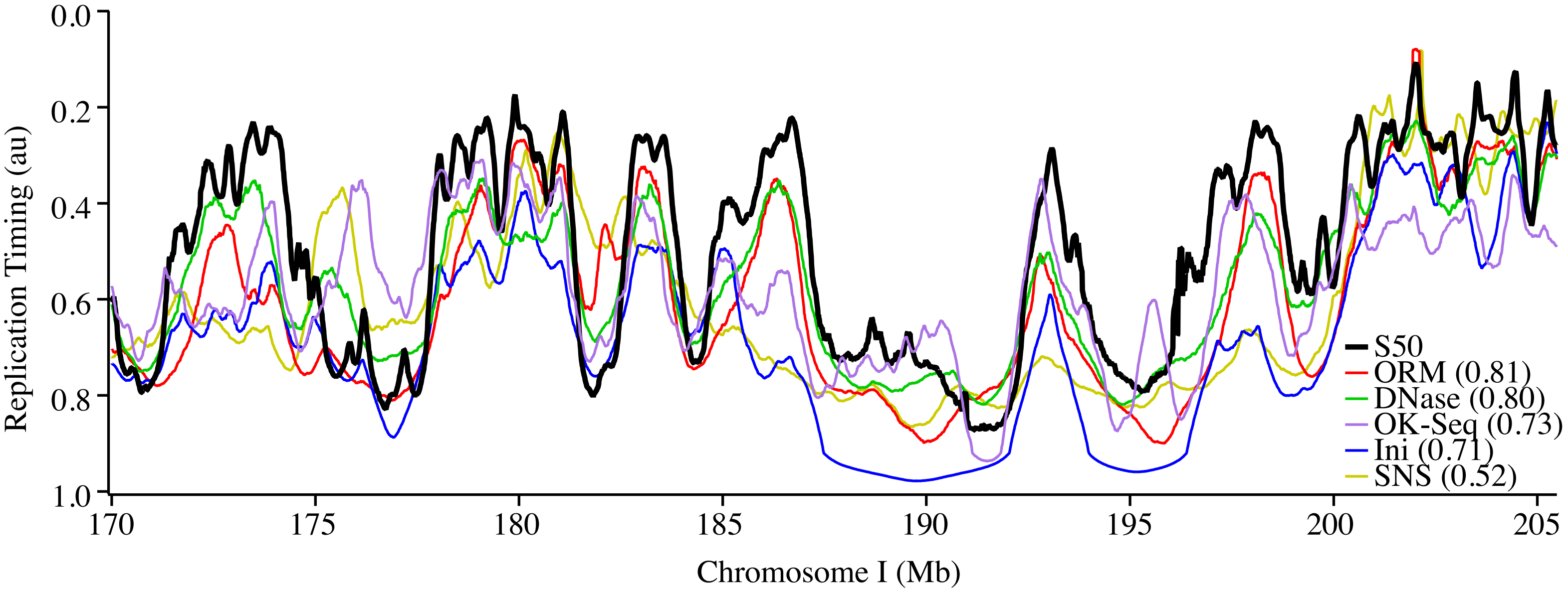

Figure S5. Simulated Replication Timing Profiles
Comparison between experimentally determined replication timing (S50 from Chen et al., 2010) and replication timing predicted from ORM, DNase I hypersensitivity (Bernstein et al., 2012), OK-Seq (Petryk et al., 2016), Ini-seq (Langley et al., 2016) and SNS-seq (Picard et al., 2014) data using a stochastic model (Gindin et al., 2014a).

Supplemental Mathematical Methods

1 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 λ_p. 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 Λ_f. Therefore, the distribution of the number of photons is given by

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

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

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

where c_0 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 $\exp(-\Lambda_f + c_0)$, to get

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

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

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

2 Probability Distribution of Intersignal Distances

We seek the intersignal distance distribution, $p_\ell(\ell)$. 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 $p(x, x+\ell)$ is the joint probability to find one signal at position x and another at $x+\ell$, without any signal in between them. $p_\ell(\ell)$ 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 decreases by a factor $\mathrm{e}^{-1} \approx 0.37$ 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.

3 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 $x_1, x_2, ..., x_n$ (we set $x_1 < x_2 < ... < n_n$). If we assume that the segment was initiated at position y, then the probability to observe signals at positions $x_1, x_2, ..., x_n$, 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 $\lambda(x)$ is the probability to label at a distance x from the initiation,

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

R_0 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 $x_{\frac{n}{2}}$ and $x_{\frac{n}{2}+1}$. 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 10^5 simulations using custom C code (available by request). With the correct fit parameters, this gives us a standard deviation of

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

4 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, p_c, and the probability that a signal is incorporated at the end of an IZ, p_e, are related via

$$\frac{p_e}{p_c} = \exp\left(-\frac{L}{2l}\right) , \tag{21}$$

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

$$p_e = \frac{R}{l}\int_0^L \mathrm{d}x\, \mathrm{e}^{-\frac{x}{l}} = R\left(1 - \mathrm{e}^{-\frac{L}{l}}\right) , \tag{22}$$

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

$$p_c = \frac{R}{l}\int_0^L \mathrm{d}x\, \mathrm{e}^{-\frac{\left|x-\frac{L}{2}\right|}{l}} = 2R\left(1 - \mathrm{e}^{-\frac{L}{2l}}\right) , \tag{23}$$

which leads to

$$\frac{p_e}{p_c} = \frac{1 - \mathrm{e}^{-\frac{L}{l}}}{2(1 - \mathrm{e}^{-\frac{L}{2l}})} . \tag{24}$$

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 (N_e) and the number of signals within 5 kb of the center of the IZ. One then expects

$$\frac{N_e}{N_c} = \frac{p_e}{p_c} . \tag{25}$$

Therefore, we compare N_e/N_c with Eqs. (21) and (23), where we have determined l from the intersignal distance, $l = 10^5$.

5 Correlation of Labeling in Neighboring Initiation Zones

We consider the correlation function that two neighboring IZ's, i and $i+1$, are labelled,

$$C = \frac{\langle a_i a_{i+1}\rangle - \langle a_i\rangle \langle a_{i+1}\rangle}{\sigma_{a_i}\sigma_{a_{i+1}}} , \tag{26}$$

where $a_i = 1$ if IZ i is labeled and 0 if not, and σ_{a_i} is the standard deviation of a_i. As a_i is a binary observable, one can easily check that

$$\sigma_{a_i} = \sqrt{\langle a_i \rangle \left(1 - \langle a_i \rangle\right)}\,, \tag{27}$$

and similarly for a_{i+1}.

Let us assume a model of uncorrelated initiations and, for the moment, that fluctuations in amount of label are irrelevant. In this model, the only correlation between a_i and a_{i+1} comes from passive replication: if IZ i initiates, there is a measurable chance that IZ $i + 1$ gets labelled because it is passively replicated by the fork initiated in IZ i. As the amount of label, and therefore the probability to get labelled, decays exponentially with the distance between the IZs, with rate the depletion rate, one would expect that, approximately,

$$\langle a_i a_{i+1} \rangle = \langle a_i \rangle \langle a_{i+1} \rangle + c \exp\left(-\frac{x}{\ell_0}\right), \tag{28}$$

where c is an unknown parameter, x is the distance between i and $i + 1$ and ℓ_0 is the (1/e) depletion length scale (106 ± 1 kb, Figure S2B).

Let us now include the effect of fluctuations in amount of label between cells. One can write

$$\langle a_i a_{i+1} \rangle = \left\langle \frac{R^2}{R_0^2} a_{0;i} a_{0;i+1} \right\rangle = \frac{\left\langle R^2 \right\rangle \langle a_{0;i} a_{0;i+1} \rangle}{R_0^2}\,, \tag{29}$$

where R is the amount of label in the cell, R_0 is the average amount of label in a cell and $a_{0;i}$ is a_i if the amount of label were to be the average amount of label. If we plug in Eqs. 27, 28 and 29 into Eq. 26, we get

$$C = \frac{c_1 \mathrm{e}^{-\frac{x}{\ell_0}} + c_2 \langle a_i \rangle \langle a_{i+1} \rangle}{\sqrt{\langle a_i \rangle \left(1 - \langle a_i \rangle\right) \langle a_{i+1} \rangle \left(1 - \langle a_{i+1} \rangle\right)}}\,, \tag{30}$$

where

$$c_1 = c\frac{\left\langle R^2 \right\rangle}{R_0^2}, \qquad c_2 = \frac{\left\langle R^2 \right\rangle - R_0^2}{R_0^2}\,. \tag{31}$$

are fit parameters. We leave c_1 as a fit parameter, and from the experimental data, we estimate $c_2 = 0.55 \pm 0.05$ (Figure S1B). Fitting gives

$$c_1 = 0.016 \pm 0.002\,. \tag{32}$$

References

[1] M. Abramowitz, I. Stegun, Eds., *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, US Government printing office (1948).

References

Abdurashidova, G., Deganuto, M., Klima, R., Riva, S., Biamonti, G., Giacca, M., and Falaschi, A. (2000). Start sites of bidirectional DNA synthesis at the human lamin B2 origin. Science *287*, 2023-2026.

Anglana, M., Apiou, F., Bensimon, A., and Debatisse, M. (2003). Dynamics of DNA replication in mammalian somatic cells: nucleotide pool modulates origin choice and interorigin spacing. Cell *114*, 385-394.

Bechhoefer, J., and Rhind, N. (2012). Replication timing and its emergence from stochastic processes. Trends Genet *28*, 374-381.

Berezney, R., Dubey, D. D., and Huberman, J. A. (2000). Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. Chromosoma *108*, 471-484.

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Besnard, E., Babled, A., Lapasset, L., Milhavet, O., Parrinello, H., Dantec, C., Marin, J. M., and Lemaitre, J. M. (2012). Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. Nat Struct Mol Biol

Birnbaum, A. (1954). Statistical Methods for Poisson Processes and Exponential Populations. Journal of the American Statistical Association *49*, 254-266.

Blow, J. J., Gillespie, P. J., Francis, D., and 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.

Cayrou, C., Ballester, B., Peiffer, I., Fenouil, R., Coulombe, P., Andrau, J. C., van Helden, J., and Mechali, M. (2015). The chromatin environment shapes DNA replication origin organization and defines origin classes. Genome Res *25*, 1873-1885.

Cayrou, C., Coulombe, P., Vigneron, A., Stanojcic, S., Ganier, O., Peiffer, I., Rivals, E., Puy, A., Laurent-Chabalier, S., Desprat, R., and Mechali, 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.

Chagin, V. O., Casas-Delucchi, C. S., Reinhart, M., Schermelleh, L., Markaki, Y., Maiser, A., Bolius, J. J., Bensimon, A., Fillies, M., Domaing, P., Rozanov, Y. M., Leonhardt, H., and Cardoso, M. C. (2016). 4D Visualization of replication foci in mammalian cells corresponding to individual replicons. Nat Commun *7*, 11231.

Chen, C. L., Duquenne, L., Audit, B., Guilbaud, G., Rappailles, A., Baker, A., Huvet, M., d'Aubenton-Carafa, Y., Hyrien, O., Arneodo, A., and Thermes, C. (2011). Replication-associated mutational asymmetry in the human genome. Mol Biol Evol *28*, 2327-2337.

Chen, C. L., Rappailles, A., Duquenne, L., Huvet, M., Guilbaud, G., Farinelli, L., Audit, B., d'Aubenton-Carafa, Y., Arneodo, A., Hyrien, O., and Thermes, C. (2010). Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res *20*, 447-457.

Conti, C., Saccà, B., Herrick, J., Lalou, C., Pommier, Y., and Bensimon, A. (2007). Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. Mol Biol Cell *18*, 3059-3067.

Czajkowsky, D. M., Liu, J., Hamlin, J. L., and Shao, Z. (2008). DNA Combing Reveals Intrinsic Temporal Disorder in the Replication of Yeast Chromosome VI. J Mol Biol *375*, 12-19.

De Carli, F., Menezes, N., Berrabah, W., Barbe, V., Genovesio, A., and Hyrien, O. (2018). High-Throughput Optical Mapping of Replicating DNA. Small Methods *0*, 1800146.

de Moura, A. P., Retkute, R., Hawkins, M., and Nieduszynski, C. A. (2010). Mathematical modelling of whole chromosome replication. Nucleic Acids Res *38*, 5623-5633.

Dellino, G. I., Cittaro, D., Piccioni, R., Luzi, L., Banfi, S., Segalla, S., Cesaroni, M., Mendoza-Maldonado, R., Giacca, M., and 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.

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

Dijkwel, P. A., Wang, S., and 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.

Dileep, V., and Gilbert, D. M. (2018). Single-cell replication profiling to measure stochastic variation in mammalian replication timing. Nat Commun *9*, 427.

Eaton, M. L., Galani, K., Kang, S., Bell, S. P., and Macalpine, D. M. (2010). Conserved nucleosome positioning defines replication origins. Genes Dev *24*, 748-753.

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

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin-state discovery and characterization. Nat Methods *9*, 215-216.

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

Feijoo, C., Hall-Jackson, C., Wu, R., Jenkins, D., Leitch, J., Gilbert, D. M., and Smythe, C. (2001). Activation of mammalian Chk1 during DNA replication arrest: a role for Chk1 in the intra-S phase checkpoint monitoring replication origin firing. J Cell Biol *154*, 913-923.

Foulk, M. S., Urban, J. M., Casella, C., and Gerbi, S. A. (2015). Characterizing and controlling intrinsic biases of lambda exonuclease in nascent strand sequencing reveals phasing between nucleosomes and G-quadruplex motifs around a subset of human replication origins. Genome Res *25*, 725-735.

Fragkos, M., Ganier, O., Coulombe, P., and Méchali, M. (2015). DNA replication origin activation in space and time. Nat Rev Mol Cell Biol *16*, 360-374.

Ganier, O., Prorok, P., Akerman, I., and Méchali, M. (2019). Metazoan DNA replication origins. Curr Opin Cell Biol *58*, 134-141.

Georgieva, D., Liu, Q., Wang, K., and Egli, D. (2019). Detection of Base Analogs Incorporated During DNA Replication by Nanopore Sequencing. bioRxiv 549220.

Gindin, Y., Meltzer, P. S., and Bilke, S. (2014a). Replicon: a software to accurately predict DNA replication timing in metazoan cells. Front Genet *5*, 378.

Gindin, Y., Valenzuela, M. S., Aladjem, M. I., Meltzer, P. S., and Bilke, S. (2014b). A chromatin structure-based model accurately predicts DNA replication timing in human cells. Mol Syst Biol *10*, 722.

Goldar, A., Marsolier-Kergoat, M. C., and Hyrien, O. (2009). Universal temporal profile of replication origin activation in eukaryotes. PLoS One *4*, e5899.

Gros, J., Kumar, C., Lynch, G., Yadav, T., Whitehouse, I., and Remus, D. (2015). Post-licensing Specification of Eukaryotic Replication Origins by Facilitated Mcm2-7 Sliding along DNA. Mol Cell *60*, 797-807.

Guilbaud, G., Rappailles, A., Baker, A., Chen, C. L., Arneodo, A., Goldar, A., d'Aubenton-Carafa, Y., Thermes, C., Audit, B., and Hyrien, O. (2011). Evidence for Sequential and Increasing Activation of Replication Origins along Replication Timing Gradients in the Human Genome. PLoS Comput Biol *7*, e1002322.

Hamlin, J. L., Mesner, L. D., Lar, O., Torres, R., Chodaparambil, S. V., and Wang, L. (2008). A revisionist replicon model for higher eukaryotic genomes. J Cell Biochem *105*, 321-329.

Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., Weaver, M., Dorschner, M. O., Gartler, S. M., and Stamatoyannopoulos, J. A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci U S A *107*, 139-144.

Hennion, M., Arbona, J. M., Lacroix, L., Cruaud, C., Theulot, B., Tallec, B. L., Proux, F., Wu, X., Novikova, E., Engelen, S., Lemainque, A., Audit, B., and Hyrien, O. (2020). FORK-seq: replication landscape of the Saccharomyces cerevisiae genome by nanopore sequencing. Genome Biol *21*, 125.

Herrick, J., and Bensimon, A. (1999). Single molecule analysis of DNA replication. Biochimie *81*, 859-871.

Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., Chang, C. W., Lyou, Y., Townes, T. M., Schubeler, D., and Gilbert, D. M. (2008). Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol *6*, e245.

Huberman, J. A., and Riggs, A. D. (1968). On the mechanism of DNA replication in mammalian chromosomes. J Mol Biol *32*, 327-341.

Huberman, J. A., and Tsai, A. (1973). Direction of DNA replication in mammalian cells. J Mol Biol *75*, 5-12.

Hyrien, O. (2015). Peaks cloaked in the mist: the landscape of mammalian replication origins. J Cell Biol *208*, 147-160.

Jackson, D. A., and 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.

Kaykov, A., and Nurse, P. (2015). The spatial and temporal organization of origin firing during the S-phase of fission yeast. Genome Res *25*, 391-401.

Kaykov, A., Taillefumier, T., Bensimon, A., and Nurse, P. (2016). Molecular Combing of Single DNA Molecules on the 10 Megabase Scale. Sci Rep *6*, 19636.

Keller, C., Ladenburger, E. M., Kremer, M., and Knippers, R. (2002). The origin recognition complex marks a replication origin in the human TOP1 gene promoter. J Biol Chem *277*, 31430-31440.

Lam, E. T., Hastie, A., Lin, C., Ehrlich, D., Das, S. K., Austin, M. D., Deshpande, P., Cao, H., Nagarajan, N., Xiao, M., and Kwok, P. Y. (2012). Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol *30*, 771-776.

Langley, A. R., Gräf, S., Smith, J. C., and 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.

Lebofsky, R., Heilig, R., Sonnleitner, M., Weissenbach, J., and Bensimon, A. (2006). DNA Replication Origin Interference Increases the Spacing between Initiation Events in Human Cells. Mol Biol Cell *17*, 5337-5345.

Löb, D., Lengert, N., Chagin, V. O., Reinhart, M., Casas-Delucchi, C. S., Cardoso, M. C., and Drossel, B. (2016). 3D replicon distributions arise from stochastic initiation and domino-like DNA replication progression. Nat Commun *7*, 11207.

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., and Li, G. (2020). H2A.Z facilitates licensing and activation of early replication origins. Nature *577*, 576-581.

Lubelsky, Y., Sasaki, T., Kuipers, M. A., Lucas, I., Le Beau, M. M., Carignon, S., Debatisse, M., Prinz, J. A., Dennis, J. H., and Gilbert, D. M. (2011). Pre-replication complex proteins assemble at regions of low nucleosome occupancy within the Chinese hamster dihydrofolate reductase initiation zone. Nucleic Acids Res *39*, 3141-3155.

Macheret, M., and Halazonetis, T. D. (2018). Intragenic origins due to short G1 phases underlie oncogene-induced DNA replication stress. Nature *555*, 112-116.

Marchal, C., Sima, J., and Gilbert, D. M. (2019). Control of DNA replication timing in the 3D genome. Nat Rev Mol Cell Biol *20*, 721-737.

Marheineke, K., and 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.

Masai, H., Matsumoto, S., You, Z., Yoshizawa-Sugata, N., and Oda, M. (2010). Eukaryotic chromosome DNA replication: where, when, and how? Annu Rev Biochem *79*, 89-130.

McGuffee, S. R., Smith, D. J., and Whitehouse, I. (2013). Quantitative, genome-wide analysis of eukaryotic replication initiation and termination. Mol Cell *50*, 123-135.

Mesner, L. D., Li, X., Dijkwel, P. A., and Hamlin, J. L. (2003). The dihydrofolate reductase origin of replication does not contain any nonredundant genetic elements required for origin activity. Mol Cell Biol *23*, 804-814.

Mesner, L. D., Valsakumar, V., Cieslik, M., Pickin, R., Hamlin, J. L., and 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.

Miotto, B., Ji, Z., and Struhl, K. (2016). Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. Proc Natl Acad Sci U S A *113*, E4810-9.

Müller, C. A., Boemo, M. A., Spingardi, P., Kessler, B. M., Kriaucionis, S., Simpson, J. T., and Nieduszynski, C. A. (2019). Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. Nat Methods 442814.

Norio, P., and Schildkraut, C. L. (2001). Visualization of DNA replication on individual Epstein-Barr virus episomes. Science *294*, 2361-2364.

O'Keefe, R. T., Henderson, S. C., and Spector, D. L. (1992). Dynamic organization of DNA replication in mammalian cell nuclei: spatially and temporally defined replication of chromosome-specific alpha-satellite DNA sequences. J Cell Biol *116*, 1095-1110.

Panning, M. M., and Gilbert, D. M. (2005). Spatio-temporal organization of DNA replication in murine embryonic stem, primary, and immortalized cells. J Cell Biochem *95*, 74-82.

Patel, P. K., Arcangioli, B., Baker, S. P., Bensimon, A., and Rhind, N. (2006). DNA replication origins fire stochastically in fission yeast. Mol Biol Cell *17*, 308-316.

Petryk, N., Kahli, M., d'Aubenton-Carafa, Y., Jaszczyszyn, Y., Shen, Y., Silvain, M., Thermes, C., Chen, C. L., and Hyrien, O. (2016). Replication landscape of the human genome. Nat Commun *7*, 10208.

Picard, F., Cadoret, J. C., Audit, B., Arneodo, A., Alberti, A., Battail, C., Duret, L., and Prioleau, M. N. (2014). The spatiotemporal program of DNA replication is associated with specific combinations of chromatin marks in human cells. PLoS Genet *10*, e1004282.

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., Thurman, R. E., Cheng, Y., Gulsoy, G., Dennis, J. H., Snyder, M. P., Stamatoyannopoulos, J. A., Taylor, J., Hardison, R. C., Kahveci, T., Ren, B., and Gilbert, D. M. (2014). Topologically associating domains are stable units of replication-timing regulation. Nature *515*, 402-405.

Pourkarimi, E., Bellush, J. M., and Whitehouse, I. (2016). Spatiotemporal coupling and decoupling of gene transcription with DNA replication origins during embryogenesis in C. elegans. Elife *5*,

Powell, S. K., MacAlpine, H. K., Prinz, J. A., Li, Y., Belsky, J. A., and MacAlpine, D. M. (2015). Dynamic loading and redistribution of the Mcm2-7 helicase complex through the cell cycle. EMBO J *34*, 531-543.

Prorok, P., Artufel, M., Aze, A., Coulombe, P., Peiffer, I., Lacroix, L., Guédin, A., Mergny, J. L., Damaschke, J., Schepers, A., Cayrou, C., Teulade-Fichou, M. P., Ballester, B., and Méchali, M. (2019). Involvement of G-quadruplex regions in mammalian replication origin activity. Nat Commun *10*, 3274.

Puig Lombardi, E., Holmes, A., Verga, D., Teulade-Fichou, M. P., Nicolas, A., and Londoño-Vallejo, A. (2019). Thermodynamically stable and genetically unstable G-quadruplexes are depleted in genomes across species. Nucleic Acids Res *47*, 6098-6113.

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., Heyne, S., Dündar, F., and Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res *44*, W160-5.

Rhind, N., and Gilbert, D. M. (2013). DNA replication timing. Cold Spring Harb Perspect Biol *5*, a010132.

Rhind, N., Yang, S. C.-H., and Bechhoefer, J. (2010). Reconciling stochastic origin firing with defined replication timing. Chromosome Res *18*, 35-43.

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011). Integrative genomics viewer. Nat Biotechnol *29*, 24-26.

Tao, L., Dong, Z., Leffak, M., Zannis-Hadjopoulos, M., and Price, G. (2000). Major DNA replication initiation sites in the c-myc locus in human cells. J Cell Biochem *78*, 442-457.

Taylor, J. H. (1977). Increase in DNA replication sites in cells held at the beginning of S phase. Chromosoma *62*, 291-300.

Técher, H., Koundrioukoff, S., Azar, D., Wilhelm, T., Carignon, S., Brison, O., Debatisse, M., and Le Tallec, B. (2013). Replication dynamics: biases and robustness of DNA fiber analysis. J Mol Biol *425*, 4845-4855.

Theis, J. F., and Newlon, C. S. (1997). The ARS309 chromosomal replicator of Saccharomyces cerevisiae depends on an exceptional ARS consensus sequence. Proc Natl Acad Sci U S A *94*, 10786-10791.

Valton, A. L., Hassan-Zadeh, V., Lema, I., Boggetto, N., Alberti, P., Saintome, C., Riou, J. F., and Prioleau, M. N. (2014). G4 motifs affect origin positioning and efficiency in two vertebrate replicators. EMBO J *33*, 732-746.

Vashee, S., Cvetic, C., Lu, W., Simancek, P., Kelly, T. J., and Walter, J. C. (2003). Sequence-independent DNA binding and replication initiation by the human origin recognition complex. Genes Dev *17*, 1894-1908.

Wilson, K. A., Elefanty, A. G., Stanley, E. G., and Gilbert, D. M. (2016). Spatio-temporal re-organization of replication foci accompanies replication domain consolidation during human pluripotent stem cell lineage specification. Cell Cycle *15*, 2464-2475.

Xu, J., Yanagisawa, Y., Tsankov, A. M., Hart, C., Aoki, K., Kommajosyula, N., Steinmann, K. E., Bochicchio, J., Russ, C., Regev, A., Rando, O. J., Nusbaum, C., Niki, H., Milos, P., Weng, Z., and Rhind, N. (2012). Genome-wide identification and characterization of replication origins by deep sequencing. Genome Biol *13*, R27.

Yang, S. C.-H., and Bechhoefer, J. (2008). How Xenopus laevis embryos replicate reliably: Investigating the random-completion problem. Phys Rev E Stat Nonlin Soft Matter Phys *78*, 041917.

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

Zhao, P. A., Sasaki, T., and Gilbert, D. M. (2020). High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. Genome Biol *21*, 76.