Supplemental Information

Figures S1-6, Supplemental Figures 2

Method S1, Supplemental Mathematical Methods 12

Figure S1

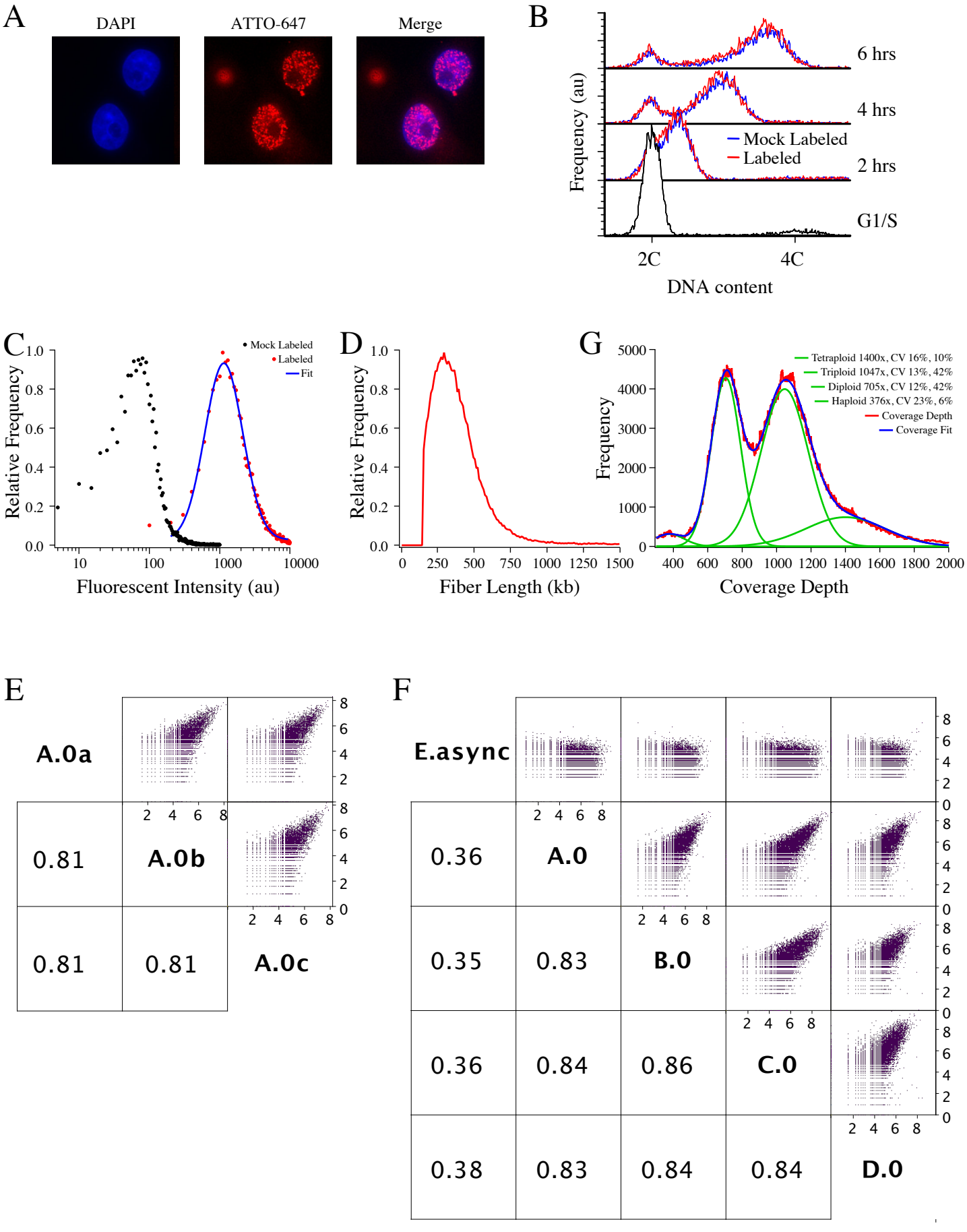
A
DAPI
ATTO-647
Merge
B
6 hrs
4 hrs
Mock Labeled
Labeled
2 hrs
G1/S
Frequency (au)
2C
4C
DNA content
C
Mock Labeled
Labeled
Fit
Relative Frequency
Fluorescent Intensity (au)
D
Relative Frequency
Fiber Length (kb)
G
Tetraploid 1400x, CV 16%, 10%
Triploid 1047x, CV 13%, 42%
Diploid 705x, CV 12%, 42%
Haploid 376x, CV 23%, 6%
Coverage Depth
Coverage Fit
Frequency
Coverage Depth
E
A.0a
0.81
A.0b
0.81
0.81
A.0c
F
E.async
0.36
A.0
0.35
0.83
B.0
0.36
0.84
0.86
C.0
0.38
0.83
0.84
0.84
D.0

Figure S1. Characterization of ORM Labeling, Related to STAR Methods

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 S-phase progression after ATTO-647-dUTP electroporation.

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

D) 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.

E) 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.

F) 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.

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

Figure S2

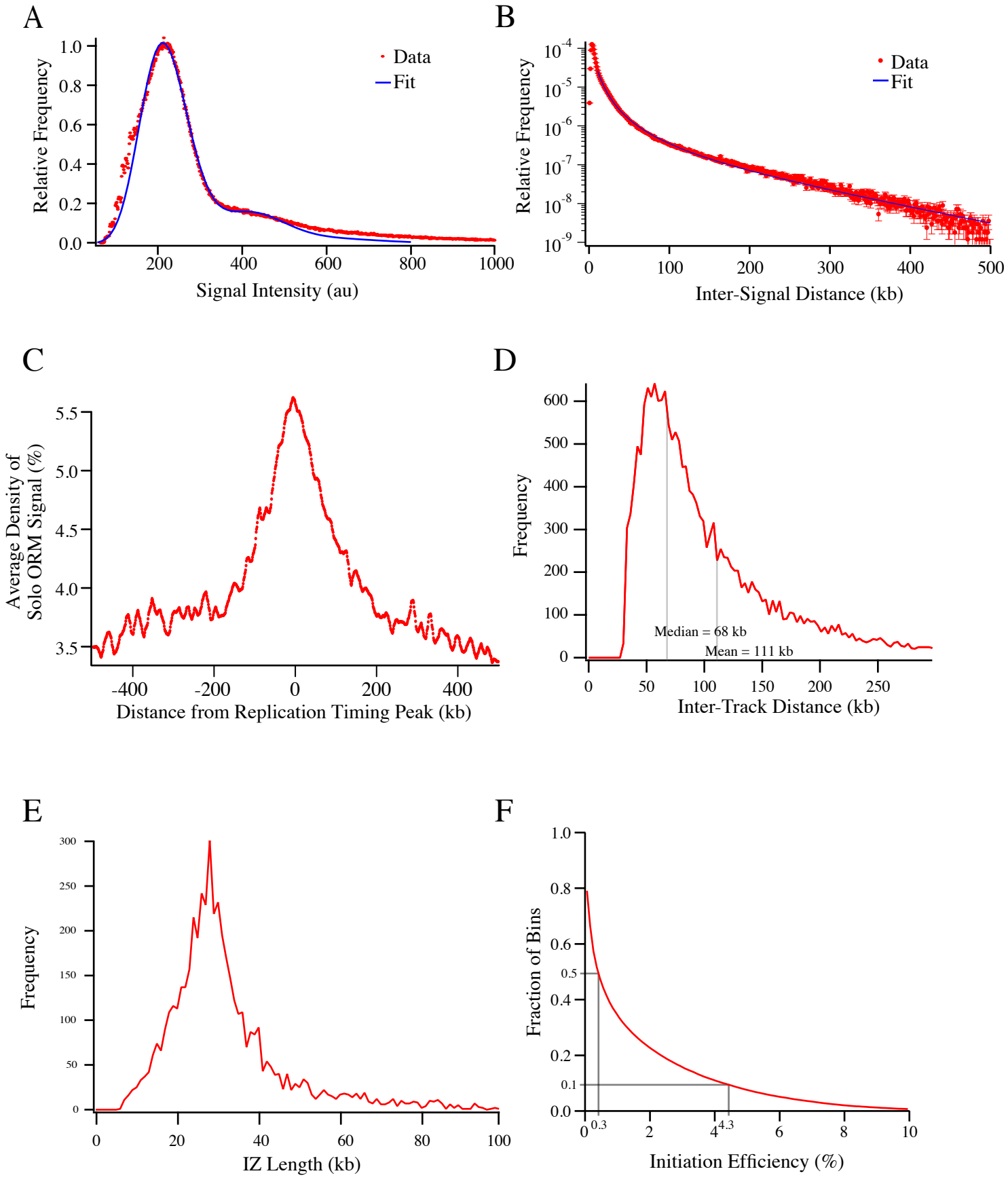
A
Relative Frequency
Signal Intensity (au)
Data
Fit
B
Relative Frequency
Inter-Signal Distance (kb)
Data
Fit
C
Average Density of Solo ORM Signal (%)
Distance from Replication Timing Peak (kb)
D
Frequency
Median = 68 kb
Mean = 111 kb
Inter-Track Distance (kb)
E
Frequency
IZ Length (kb)
F
Fraction of Bins
Initiation Efficiency (%)

Figure S2. Characterization of ORM Label Distribution, Related to STAR Methods

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, Method S1) 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, Method S1) 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 IZ lengths.

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

Figure S3

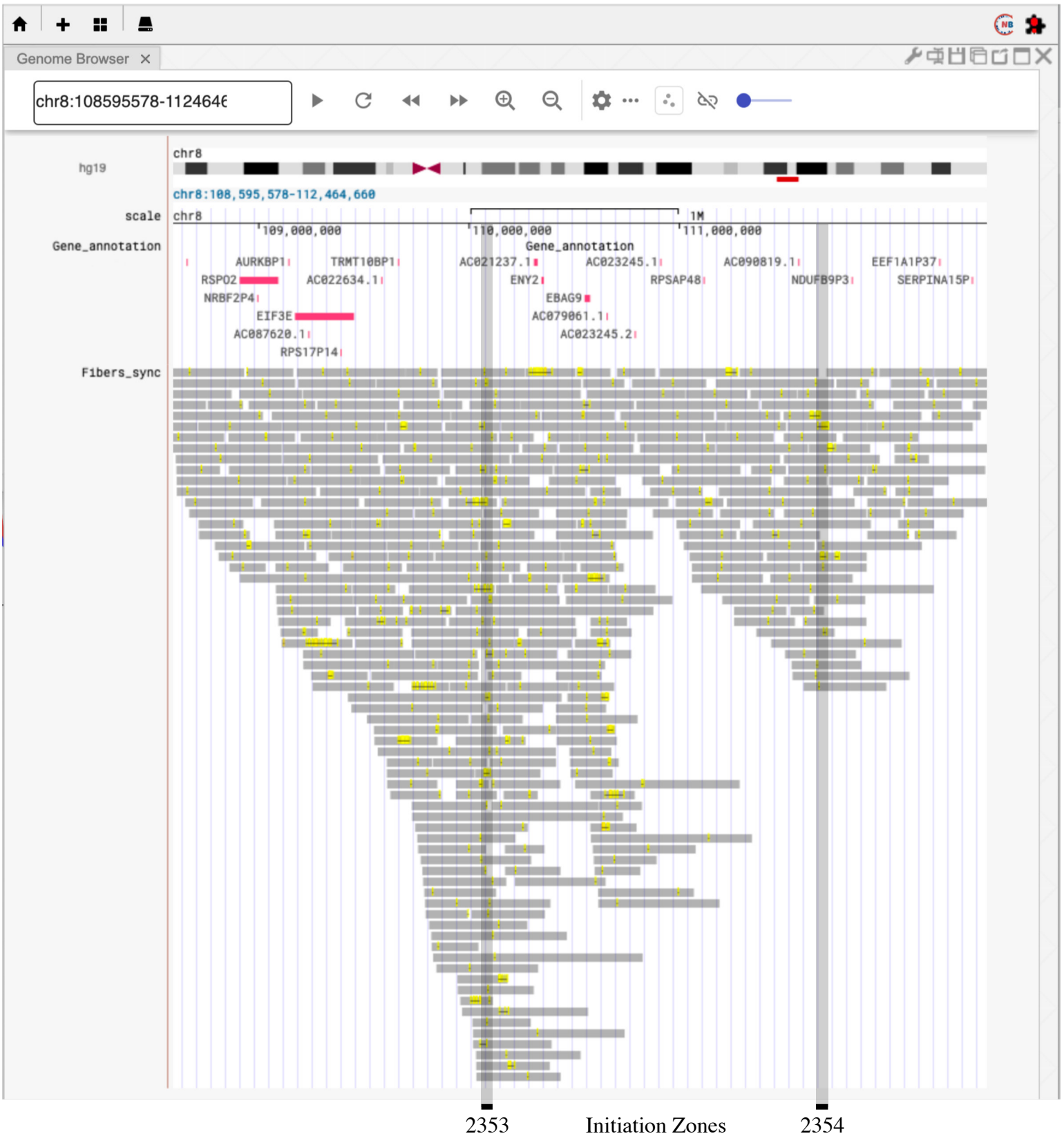

Figure S3. The ORM Genome Browser, Related to Figure 1

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

Figure S4

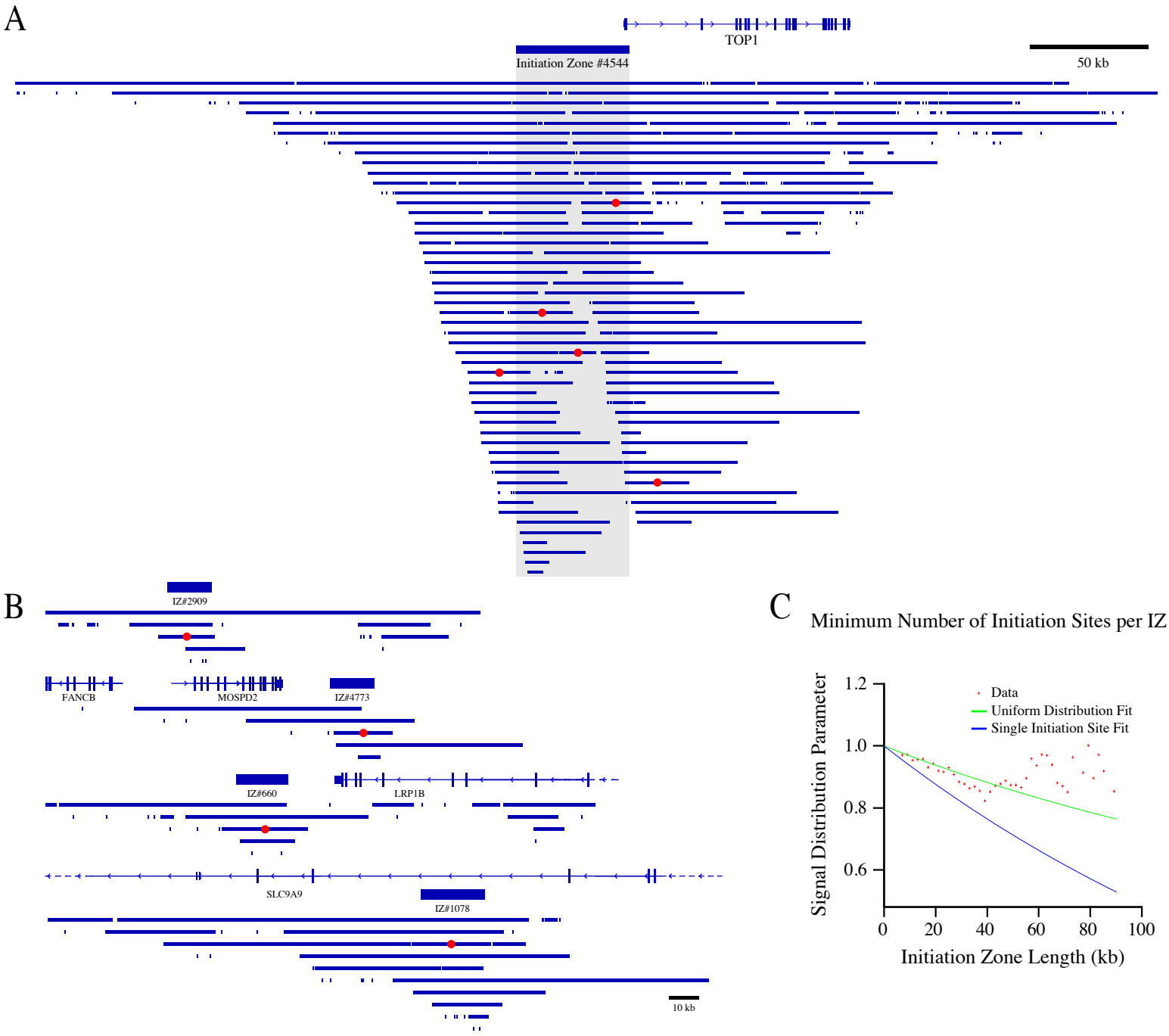

Figure S4. Distribution of Replication Tracks within Initiation Zones, Related to Figure 3

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

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

C) The distribution of signal across IZs. The ratio of signal frequency at the IZ center to the IZ boundary is plotted versus IZ length. This value is expected to decrease more quickly in IZs that predominantly have a single initiation site (Eq. 25) than if initiation is distributed across the IZ (Eq. 28, Supplemental Mathematical Methods). The distribution across IZs shorter than 55 kb is consistent with a uniform distribution of initiation sites. At longer lengths, we actually see more signal at the edges of the IZs than it the center. One possible explanation for this phenomenon is that larger IZs may actually be two smaller IZs fused together such that there is more initiation sites towards the edges of the fused IZ and less initiation in the center, which is actually between the two constituent IZs. See Method S1 for details.

Figure S5

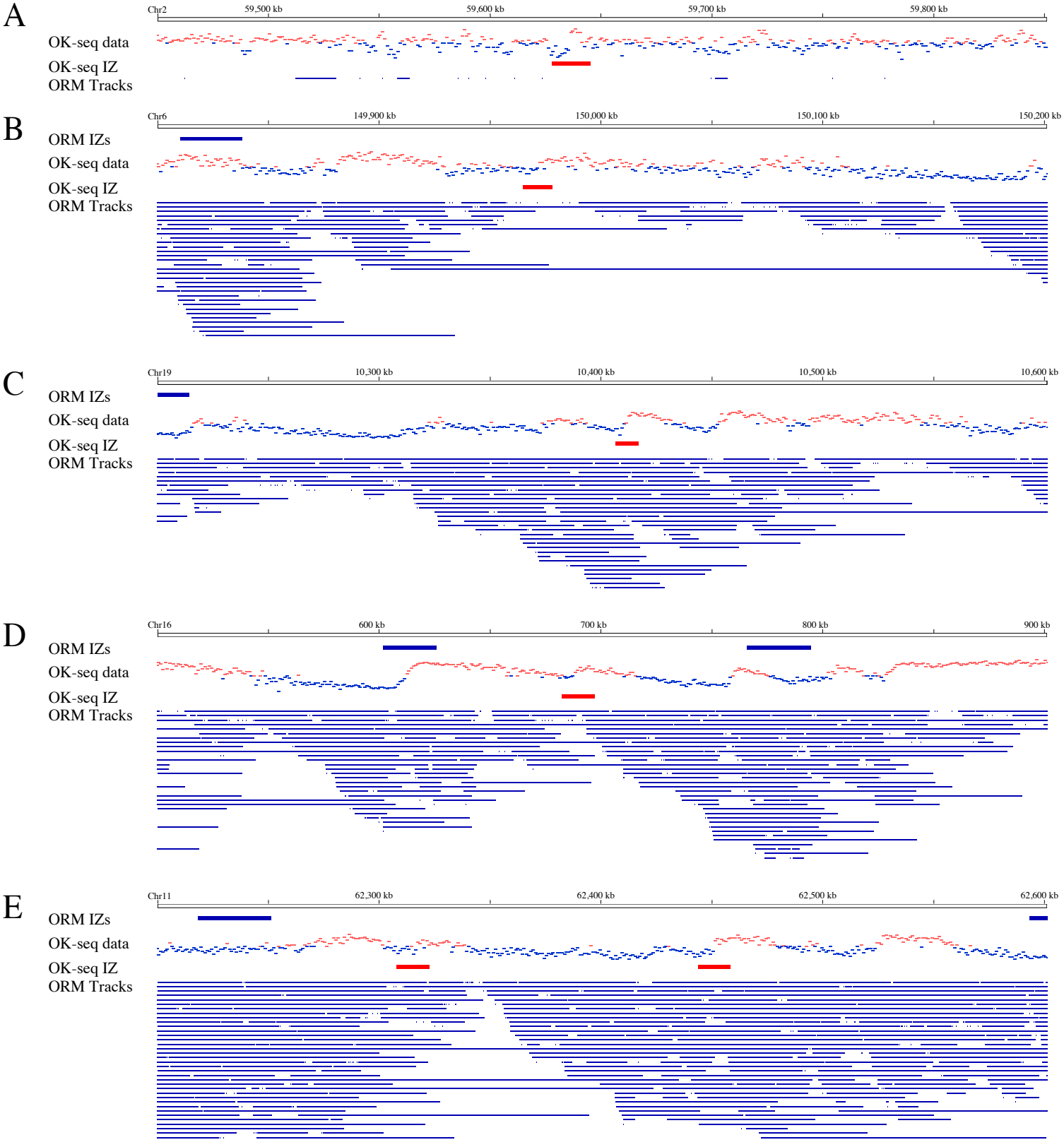
A
OK-seq data
OK-seq IZ
ORM Tracks
Chr2
59,500 kb
59,600 kb
59,700 kb
59,800 kb
B
ORM IZs
OK-seq data
OK-seq IZ
ORM Tracks
Chr6
149,900 kb
150,000 kb
150,100 kb
150,200 kb
C
ORM IZs
OK-seq data
OK-seq IZ
ORM Tracks
Chr19
10,300 kb
10,400 kb
10,500 kb
10,600 kb
D
ORM IZs
OK-seq data
OK-seq IZ
ORM Tracks
Chr16
600 kb
700 kb
800 kb
900 kb
E
ORM IZs
OK-seq data
OK-seq IZ
ORM Tracks
Chr11
62,300 kb
62,400 kb
62,500 kb
62,600 kb

Figure S5. Reanalysis of Potentially Discrete OK-Seq Initiation Zones, Related to Figure 3

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

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

C) 6 are robust transitions that correlate with numerous ORM segments. We conclude that they are IZs in both the OK-seq and ORM data. However, they show broadly dispersed ORM segments, therefore we do not believe they are unusually constrained IZs. Instead, we conclude that they are outliers in the OK-seq data that were identified as unusually narrow due to experimental variation. Panel C is an example of one such zone.

D,**E**) 7 are robust transitions that do not correlate with numerous ORM segments. They could be IZs present in the OK-seq HeLa cell line, but absent in the ORM HeLa cell line. Alternatively, they could be translocation break points in the OK-seq HeLa cell line relative to the hg19 reference sequence. Such breakpoints would explain both the sharpness of the transition and the absence of these putative IZs from the ORM data. Panel D is an example of one such zone. Panel E show two such zones that can be explained by an inversion between them.

Figure S6

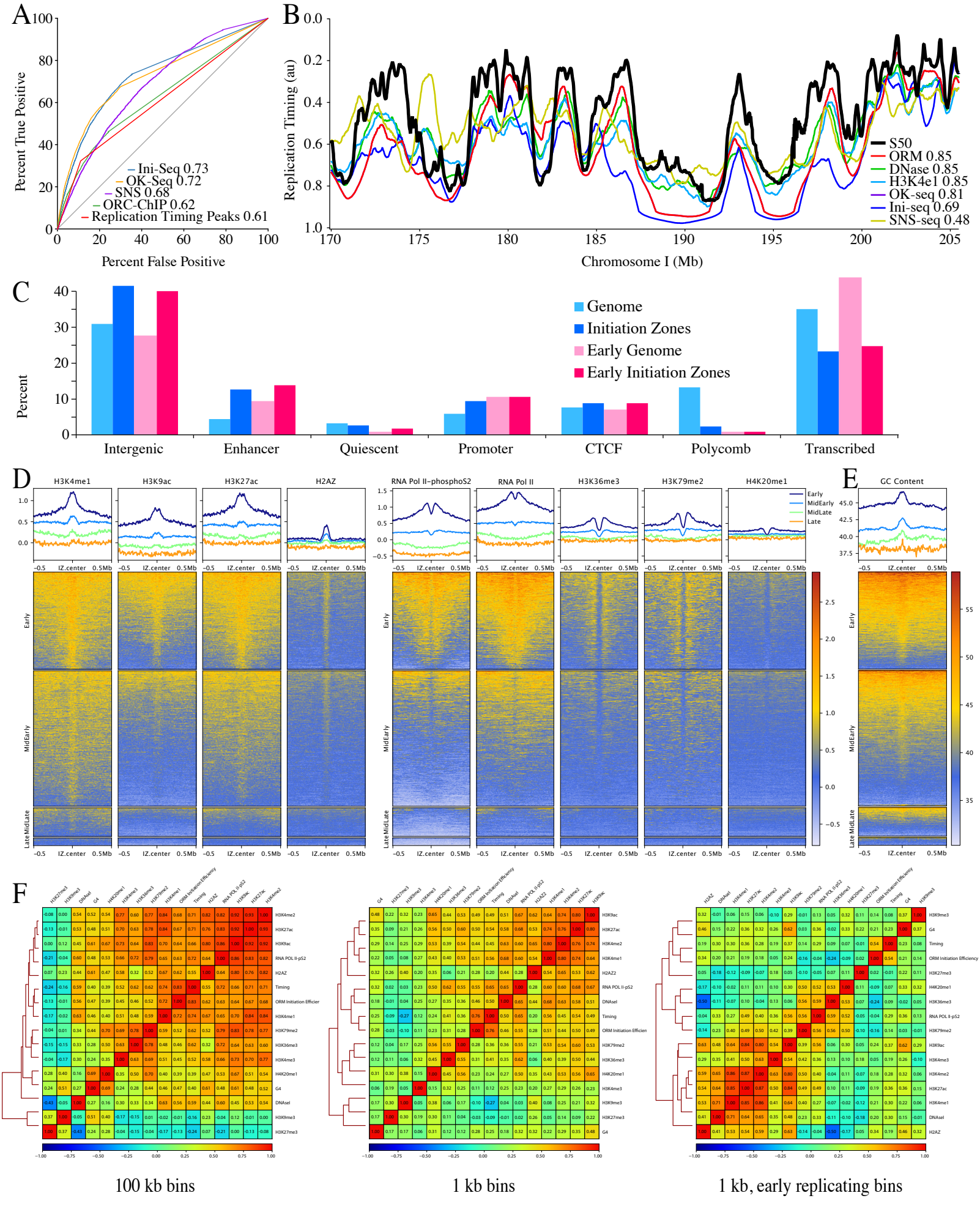
A
Percent True Positive
Percent False Positive
Ini-Seq 0.73
OK-Seq 0.72
SNS 0.68
ORC-ChIP 0.62
Replication Timing Peaks 0.61
B
Replication Timing (au)
Chromosome I (Mb)
S50
ORM 0.85
DNase 0.85
H3K4e1 0.85
OK-seq 0.81
Ini-seq 0.69
SNS-seq 0.48
C
Percent
Genome
Initiation Zones
Early Genome
Early Initiation Zones
Intergenic
Enhancer
Quiescent
Promoter
CTCF
Polycomb
Transcribed
D
H3K4me1
H3K9ac
H3K27ac
H2AZ
RNA Pol II–phosphoS2
RNA Pol II
H3K36me3
H3K79me2
H4K20me1
Early
MidEarly
MidLate
Late
IZ.center
0.5Mb
E
GC Content
F
100 kb bins
1 kb bins
1 kb, early replicating bins

Figure S6. Correlations between ORM Data and Other Datasets, Related to Figure 5, and Enrichment of Histone Modifications and Other Genomic Features in Initiation Zones, Related to Figure 4

A) ROC analysis of the association between the ORM IZs and other origin mapping data shown in Figure 3G. As shown in the ROC curves, the ORM IZs are better correlated with OK-seq IZs (AUC = 0.72) and Ini-seq IZs (AUC = 0.73). The bias of the SNS ROC curve towards high true positive rates only at high false negative values is consistent with that dataset having more false positive signal, whereas the bias of the replication timing and ORC datasets relatively high true positive rates only at low false negative values is consistent with those datasets having fewer, but more accurate true positives. Areas under the ROC curves (AUC) are shown in the legend.
B) Comparison between experimentally determined HeLa replication timing (S50) and replication timing predicted from ORM, DNase I hypersensitivity (Bernstein et al., 2012), OK-Seq (Petryk et al., 2016), Ini-seq (Langley et al., 2016) and SNS-seq (Picard et al., 2014) data using a stochastic model (Gindin et al., 2014a). The Spearman correlation coefficients with replication timing are shown in the legend.
C) 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 for all IZs and the IZs that replicate in the first quarter of S phase (S50 < 0.25).
D) The enrichment of histone modifications and other genomic features relative to ORM IZs. 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.
E) The enrichment of GC content relative to ORM IZs. The upper panels show the average % GC content signal around all IZs. The lower panels show heat maps of the % GC content at each IZ.
F) 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.

Method S1: Supplemental Mathematical Methods, Related to STAR 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],

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

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

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

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 $x_1 < x_2 < ... < x_n$). 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 Estimating the Initiation Event Labeling Efficiency

Using the analysis in Sections 1, 2 and 3, we can estimate the frequency with which an early initiation event will incorporate at least one label and thus be identified by ORM. We begin by noting that the incorporation rate of at position x of a replication that started at $t = 0$ is equal to

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

This means that at time t, this rate is

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

As $r(t)$ is independent of when the initiation started, one can see that r(x) for an initiation that started at time t_0 is given by

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

The probability to not get any signals within a distance x_0 from the initiation is then

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

Setting v=1.65 kb/min (replication fork rate, from Figure 1C), x_0=15 kb (the nominal resolution of ORM from Eq. 20), r_0=1/3.8 kb (the initial labeling rate, from Figure S2A) and c=99 kb (Figure S2b; note that the 75 kb reported there is c in base 2, whereas 99 kb used here is c in base e) and assuming that the initiations happen uniformly in early S phase, one estimates that the probability to see zero signals within the first 15 kb of an initiation is 9.5%.

5 Distribution of Signals within Initiation Zones

Consider an initiation zone of length L. We are interested in determining the distribution of initiations inside the initiation zone. Here, we will consider two

extreme cases. The first possibility is that the initiation always happens at a single point at the center of the IZ. The second possibility is that the initiation happens with equal probability everywhere along the IZ.

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

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

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

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

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

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

which leads to

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

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

$$\frac{N_e}{N_c} = \frac{p_e}{p_c}. \tag{29}$$

Therefore, we compare N_e/N_c with Eqs. (25) and (27), where we have determined l from the intersignal distance, $l = 10^5$.

6 Correlation of Labeling in Neighboring Initiation Zones

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

$$C = \frac{\langle a_i a_{i+1}\rangle - \langle a_i\rangle\langle a_{i+1}\rangle}{\sigma_{a_i}\sigma_{a_{i+1}}}, \tag{30}$$

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

$$\sigma_{a_i} = \sqrt{\langle a_i\rangle\left(1-\langle a_i\rangle\right)}, \tag{31}$$

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,

$$\langle a_i a_{i+1} \rangle = \langle a_i \rangle \langle a_{i+1} \rangle + c \exp\left(-\frac{x}{\ell_0}\right), \tag{32}$$

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

$$\langle a_i a_{i+1} \rangle = \left\langle \frac{R^2}{R_0^2} a_{0;i} a_{0;i+1} \right\rangle = \frac{\langle R^2 \rangle \langle a_{0;i} a_{0;i+1} \rangle}{R_0^2}, \tag{33}$$

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. 31, 32 and 33 into Eq. 30, we get

$$C = \frac{c_1 \mathrm{e}^{-\frac{x}{\ell_0}} + c_2 \langle a_i \rangle \langle a_{i+1} \rangle}{\sqrt{\langle a_i \rangle (1 - \langle a_i \rangle) \langle a_{i+1} \rangle (1 - \langle a_{i+1} \rangle)}}, \tag{34}$$

where

$$c_1 = c \frac{\langle R^2 \rangle}{R_0^2}, \qquad c_2 = \frac{\langle R^2 \rangle - R_0^2}{R_0^2}. \tag{35}$$

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

$$c_1 = 0.016 \pm 0.002. \tag{36}$$

References

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