



HAL
open science

Neural network and kinetic modelling of human genome replication reveal replication origin locations and strengths

Jean-Michel Arbona, Hadi Kabalane, Arach Goldar, Olivier Hyrien, Benjamin Audit

► To cite this version:

Jean-Michel Arbona, Hadi Kabalane, Arach Goldar, Olivier Hyrien, Benjamin Audit. Neural network and kinetic modelling of human genome replication reveal replication origin locations and strengths. PLoS Computational Biology, 2023, 19 (5), pp.e1011138. 10.1371/journal.pcbi.1011138 . hal-03817417

HAL Id: hal-03817417

<https://cnrs.hal.science/hal-03817417>

Submitted on 17 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 Neural network and kinetic 2 modelling of human genome 3 replication reveal replication origin 4 locations and strengths

5 Jean-Michel Arbona^{1*}, Hadi Kabalane^{2,‡}, Arach Goldar³, Olivier Hyrien^{4*},
6 Benjamin Audit^{2*}

***For correspondence:**

7 jeanmichel.arbona@ens-lyon.fr
(JMA); olivier.hyrien@bio.ens.psl.eu
(OH); benjamin.audit@ens-lyon.fr
(BA)

Present address: [‡]Centre de
Recherches en Cancérologie de
Toulouse, Inserm, Université Paul
Sabatier, CNRS, Toulouse, France.

7 ¹Laboratoire de Biologie et Modélisation de la Cellule, ENS de Lyon, Lyon, France; ²ENS
8 de Lyon, CNRS, Laboratoire de Physique, F-69342 Lyon, France; ³Ibitec-S, CEA, France;
9 ⁴Institut de Biologie de l'Ecole Normale Supérieure (IBENS), Ecole Normale Supérieure,
10 CNRS, INSERM, Université PSL, 46 rue d'Ulm, F-75005, Paris, France

12 Abstract

13 In human and other metazoans, the determinants of replication origin location and strength are
14 still elusive. Origins are licensed in G1 phase and fired in S phase of the cell cycle, respectively. It
15 is debated which of these two temporally separate steps determines origin efficiency.
16 Experiments can independently profile mean replication timing (MRT) and replication fork
17 directionality (RFD) genome-wide. Such profiles contain information on multiple origins'
18 properties and on fork speed. Due to possible origin inactivation by passive replication, however,
19 observed and intrinsic origin efficiencies can markedly differ. Thus, there is a need for methods
20 to infer intrinsic from observed origin efficiency, which is context-dependent. Here, we show that
21 MRT and RFD data are highly consistent with each other but contain information at different
22 spatial frequencies. Using neural networks, we infer an origin licensing landscape that, when
23 inserted in an appropriate simulation framework, jointly predicts MRT and RFD data with
24 unprecedented precision. We furthermore uncover an analytical formula that predicts intrinsic
25 from observed origin efficiency combined with MRT data. Comparison of inferred intrinsic origin
26 efficiencies with experimental profiles of licensed origins (ORC, MCM) and actual initiation events
27 (Bubble-seq, SNS-seq, OK-seq) show that intrinsic origin efficiency is not solely determined by
28 licensing efficiency. Thus, human replication origin efficiency is set at both the origin licensing
29 and firing steps.

31 Introduction

32 In eukaryotes, chromosome replication starts at multiple sites referred to as replication origins
33 (*DePamphilis and Bell, 2010*). Origins are licensed for replication during the G1 phase of the cell
34 cycle, when the origin recognition complex (ORC) loads the MCM2-7 replicative helicase in an inac-
35 tive, double hexameric ring form (MCM DH), around origin DNA (*Evrin et al., 2009; Remus et al.,*
36 *2009; Miller et al., 2019; Schmidt and Bleichert, 2020*). This symmetric configuration prepares the
37 helicases to initiate bidirectional replication upon activation. Origin activation (or firing) can take
38 place at different times through S phase, by binding of multiple firing factors that trigger origin
39 DNA unwinding and convert the inactive MCM DH into two active Cdc45/MCM/GINS helicases that

40 each encircles and translocates 3'-to-5' along a single DNA strand (*Douglas et al., 2018*). Only a
41 fraction of MCM DHs lead to productive initiation events, while the rest is inactivated by passing
42 replication forks originating from other origins. This origin passivation mechanism (*Arbona et al.,*
43 *2018*) cooperates with MCM2-7 loading restriction to G1 phase to prevent rereplication in a single
44 cell cycle (*Siddiqui et al., 2013*).

45 Several experimental techniques allow to monitor origin licensing and firing as well as replica-
46 tion progression during S phase. Origin licensing can be monitored by experimental detection of
47 ORC and MCM proteins, whose profiles are highly though not perfectly concordant (*Kirstein et al.,*
48 *2021; Miotto et al., 2016; Foss et al., 2021*). In contrast to these potential origin profiles, actual ini-
49 tiation events can be monitored by sequencing purified replication initiation intermediates, such
50 as short nascent DNA strands (SNS-Seq; *Picard et al. (2014)*) or bubble-containing restriction frag-
51 ments (Bubble-Seq; *Mesner et al. (2013)*). These two methods are only weakly concordant (*Hyrien,*
52 *2015; Hulke et al., 2020*). Other methods monitor replication progression along the genome. Mean
53 replication timing (MRT) profiles have been computed by sequencing newly replicated DNA from
54 sorted cells at different stages of S phase (Repli-seq; *Chen et al. (2010); Hansen et al. (2010); Zhao*
55 *et al. (2020)*) or by determining DNA copy number from proliferating cells (*Koren et al., 2014*). Peaks
56 of early MRT must contain origins, but low resolution (50-100 kb; *Chen et al. (2010); Hansen et al.*
57 *(2010); Zhao et al. (2020)*) has long precluded precise origin mapping from human MRT profiles.
58 Replication fork directionality (RFD) profiles, obtained by strand-oriented sequencing of purified
59 Okazaki fragments (OK-seq) were more resolute (< 5 kb) (*Petryk et al., 2016; Wu et al., 2018*).
60 RFD profiles revealed that: (i) each cell line contains 5,000 - 10,000 broad (10-100 kb) initiation
61 zones (IZs), characterised by a left-to-right shift in RFD; (ii) IZs often but not always flank active
62 genes; (iii) termination events occur in broad zones (TZs), characterized by a right-to-left RFD shift;
63 (iv) TZs can directly follow IZs or can be separated from IZs by extended regions of unidirectional
64 replication that lack initiation and termination events; (v) large randomly replicating regions, char-
65 acterized by extended segments of null RFD, are observed in silent heterochromatin. OK-seq IZs
66 were confirmed genome-wide by EdUseq-HU (*Tubbs et al., 2018*), high-resolution Repli-Seq (*Zhao*
67 *et al., 2020*) and Optical Replication Mapping (*Wang et al., 2021*). Importantly, initiation events may
68 additionally occur outside IZs, but in a too dispersed manner to be directly detected in cell popula-
69 tion profiles. Recent single-molecule and OK-seq analyses of the yeast genome (*Müller et al., 2019;*
70 *Hennion et al., 2020*) and of two model chicken loci (*Blin et al., 2021*) provided direct evidence for
71 dispersed initiation between efficient IZs in these two systems.

72 IZs can be shared between cell types or specific to a cell type, suggesting epigenetic regulation.
73 They are enriched in DNase I hypersensitive sites (HSSs) and histone modifications or variants such
74 as H3K4me1, H3K27ac and H2A.Z, that usually mark active transcriptional regulatory elements
75 (*Petryk et al., 2016; Wu et al., 2018; Petryk et al., 2018*). H2A.Z was proposed to facilitate origin
76 licensing and firing by recruiting SUV420H1, which promotes H4K20me2 deposition, in turn facili-
77 tating ORC binding (*Long et al., 2020*). Furthermore, binding sites for the firing factor MTBP were
78 found to colocalize with H3K4me1, H3K27ac, H2A.Z, and other active chromatin marks (*Kumagai*
79 *and Dunphy, 2020*).

80 What mechanisms could regulate origin firing? Modeling studies showed that a probabilistic
81 interaction of potential origins with rate-limiting firing factors, engaged with forks and recycled
82 at termination events, can predict the time-dependent profiles of origin firing rate and fork den-
83 sity universally observed in eukaryotes (*Arbona et al., 2018; Goldar et al., 2008, 2009*). Experi-
84 mental studies indeed suggested that rate-limiting activators regulate replication kinetics in yeast
85 (*Mantiero et al., 2011; Tanaka et al., 2011*) and metazoans (*Wong et al., 2011; Collart et al., 2013*).
86 Thus, a simple model for replication regulation is that potential origins fire at different mean times
87 because of their different affinities for limiting factors (*Douglas and Diffley, 2012*). Alternatively,
88 potential origins may all have the same affinity for firing factors but their variable density along
89 the genome may determine MRT (*Yang et al., 2010; Das et al., 2015*). We refer to these two distinct
90 models as the origin affinity model and the origin density model, respectively.

91 Modelling studies indicate that the reproducible spatial structure of genomic replication profiles
92 can emerge from stochastic firing of individual origins (*Bechhoefer and Rhind, 2012; Gindin et al.,*
93 *2014*). The latter built a kinetic model in which the time-dependent probability of initiation at a
94 yet unreplicated site was the product of the time-dependent availability of a limiting factor by the
95 time-independent, local value of a genomic “initiation probability landscape” (IPLS). Of the various
96 genomic and epigenomic profiles used as estimates of IPLS, DNase I HSS profiles produced the best
97 match with experimental MRT profiles (Pearson correlation between simulated and experimental
98 MRT of 0.865). Importantly, the same IPLSs did not produce realistic MRT profiles in models that
99 did not include competition for limiting fork-associating factors (*Gindin et al., 2014*). Since this
100 model did not explicitly separate origin licensing and firing, however, it remained unclear whether
101 the IPLS reflected potential origin density, or affinity, or both.

102 Current experimental evidence has not yet clearly distinguished between the origin affinity and
103 origin density models. ORC and MCM abundance profiles, which presumably reflect potential ori-
104 gin density, are well correlated with DNase I HSS and early MRT (*Miotto et al., 2016; Foss et al., 2021;*
105 *Kirstein et al., 2021*). Furthermore, ORC- or MCM-based IPLSs produced realistic MRT profiles in
106 Gindin-like simulations (*Miotto et al., 2016; Foss et al., 2021*), which supports the origin density
107 model. However, our comparison of ORC, MCM and RFD profiles of the Raji cell line showed that
108 when confounding parameters such as MRT and transcription status are controlled, ORC and MCM
109 densities are not predictive of IZs (*Kirstein et al., 2021*). This suggested that potential origins may
110 be more widespread than initiation sites but have different firing efficiencies, perhaps due to spe-
111 cific MCM or histone modifications affecting their affinities for firing factors, in line with the origin
112 affinity model.

113 In the present work, we harness our previous kinetic model of DNA replication (*Arbona et al.,*
114 *2018*) to predict MRT and RFD profiles. Discrete, localized potential origins (MCM DHs), chosen
115 from an arbitrary potential origin density landscape (PODLS), are activated in a stochastic manner
116 by interaction with limiting firing factors that engage with forks and are released at termination (Fig.
117 1). As each potential origin is given the same probability to fire, the non-uniformity of the obtained
118 replication profiles only comes from the non-uniformity of the PODLS (origin density model).

119 Our aim being to extract the PODLS that best predicts available MRT and RFD data, we first
120 compare their information contents. We show a remarkable conformity of MRT and RFD data to a
121 simple mathematical equation that links both profiles. Extending the work by Gindin et al. (*Gindin*
122 *et al., 2014*), we then ask whether the correlation of DNase I HSS with origin activation seen at MRT
123 resolution (50-100 kb) still holds true at RFD resolution (< 5 kb). We demonstrate that MRT and
124 RFD data provide distinct information at different scales.

125 We then train a neural network on simulated MRT and RFD profiles to infer a PODLS that jointly
126 predicts experimental MRT and RFD almost exactly, surpassing any PODLS based on DNase I HSS,
127 ORC, MCM, Bubble-seq or SNS-seq profiles. In our model, each potential origin has the same in-
128 trinsic probability of activation per unit time. The optimized PODLS, which reflects intrinsic origin
129 efficiencies, can be directly compared with ORC and MCM profiles. To compare the PODLS to actual
130 initiation events as monitored by SNS-seq, bubble-seq or OK-seq, we establish a novel mathemat-
131 ical expression that relates observed and intrinsic origin efficiencies to MRT and therefore allows
132 us to take origin passivation effects into account. The results show that the firing probability of
133 potential origins is not uniform in time and space. Our results therefore support a combined ori-
134 gin density and affinity model and provide a basis to investigate the distinct genetic and epigenetic
135 determinants of origin licensing and firing.

136 Results

137 Information complementarity between MRT and RFD profiles

138 Previous modelling works (*Gindin et al., 2014; Löb et al., 2016*) compared simulated and exper-
139 imental human MRT profiles to constrain their parameter values. RFD profiles have now been

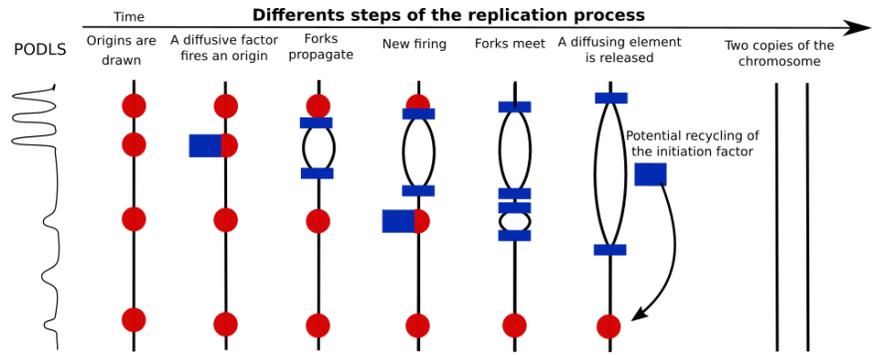


Figure 1. Modeling DNA replication. Given a PODLS derived from a specific genomic feature (e.g. a DNase I HSS profile), a fixed number of localized potential origins is drawn (red circles). Limiting firing factors (blue rectangles) activate origins in a probabilistic manner and engage with each pair of newly created forks, which propagate at velocity v . Engaged factors can no longer activate origins. Unfired origins are passivated when they are reached by a passing fork. Merging of two forks emanating from adjacent origins results in a replication termination event and the release of one firing factor which becomes available again for origin activation. MRT and RFD are then computed from the average of 200 simulations. See Materials and Methods.

140 established for many human cell lines (Petryk et al., 2016; Wu et al., 2018), providing us with an
 141 alternative comparison point. It is thus of interest to compare the information content of RFD and
 142 MRT profiles. Within the hypothesis of a constant fork speed v , MRT and RFD profiles are equiv-
 143 alent as they are analytically related to one another by (Guilbaud et al., 2011; Baker et al., 2012;
 144 Audit et al., 2012):

$$RFD(x) = v \frac{d}{dx} MRT(x). \quad (1)$$

145 Here $MRT(x)$ is the mean replication time after entry in S-phase of bin x and is expressed in time
 146 units. Note that $MRT(x)$ as measured by Repli-seq experiments is the average global replicated
 147 fraction at the moments locus x is replicated, and thus has a value between 0 and 1. Assuming
 148 a linear relation between replication fraction and S phase duration, Repli-seq experiment can be
 149 converted into time by multiplication by an estimate of S-phase duration T_S . Equation (1) can be
 150 checked using experimental data. However, the derivative of a noisy profile is ill-defined and nu-
 151 merically unstable to compute. To avoid computing the MRT derivative, Eq. (1) can be integrated
 152 at point x over a length l leading to:

$$\Delta_l MRT(x) = MRT(x+l) - MRT(x) = \frac{1}{v} \int_x^{x+l} RFD(y) dy = \frac{l}{v} \langle RFD \rangle_x^{x+l}, \quad (2)$$

153 where $\langle \cdot \rangle_x^{x+l}$ stands for the average value over $[x, x+l]$. Equation (2) predicts that the MRT
 154 change across an interval is proportional to the average RFD over that interval. Using reported
 155 Repli-seq MRT (Hansen et al., 2010) and OK-seq RFD (Wu et al., 2018) profiles for the K562 cell line,
 156 Eq. (2) was very convincingly verified over scales ranging from 10 kb to 5 Mb, with a genome-wide
 157 correlation coefficient up to 0.94 at scale 1.5 Mb, and a proportionality coefficient ranging from $v =$
 158 $1.2 \text{ kb} \cdot \text{min}^{-1}$ to $v = 1.6 \text{ kb} \cdot \text{min}^{-1}$, assuming $T_S = 12$ hours (Weis, 2012). This is illustrated on Fig. 2
 159 for scale 50 kb and Fig. S1 for other scales. Therefore, although OK-seq and Repli-seq experiments
 160 are complex and have been performed by different laboratories, they are highly consistent with
 161 each other, on a wide range of scales, within the hypothesis of a constant fork speed.

162 In their modeling work, Gindin et al (Gindin et al., 2014) found that of all epigenetic features
 163 tested, IPLSs based on DNase I HSS profiles produced the best match between simulated and
 164 experimental MRT profiles (Pearson correlation coefficient, PCC = 0.865). We performed similar
 165 simulations, using our model (Fig. 1) as detailed in Materials and Methods. Using a PODLS based
 166 on the K562 DNase I HSS profile, we drew a fixed number of potential origins and simulated a
 167 bimolecular reaction with firing factors, whose number increased from the start of S phase and

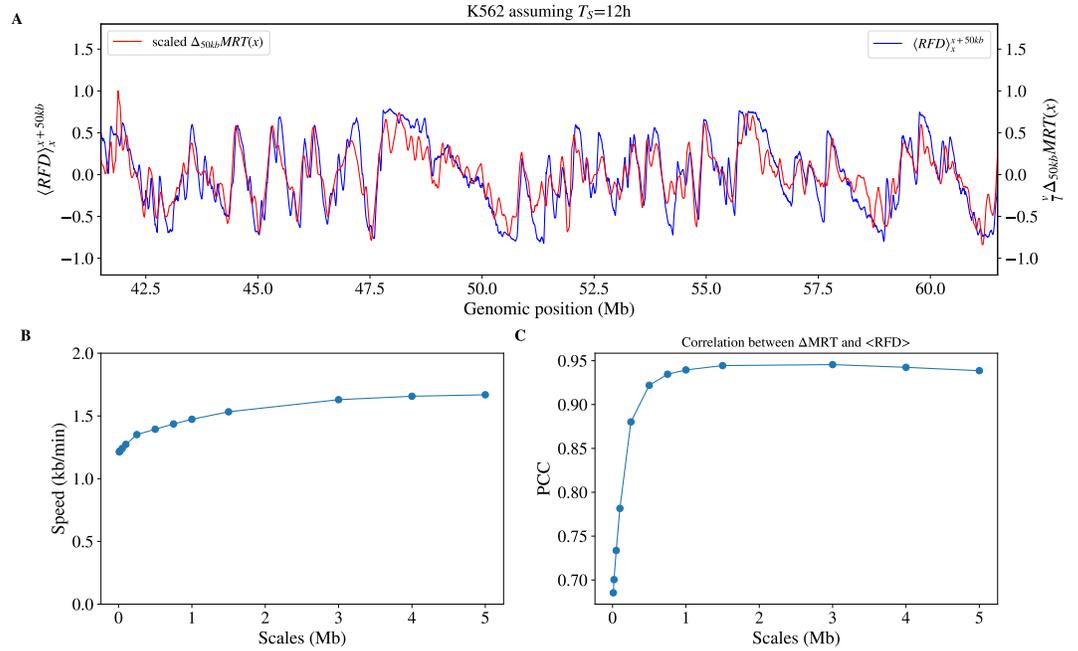


Figure 2. (A) Comparison, for a 20 Mb region of chromosome 1, of the K562 RFD profile averaged over 50 kb windows (blue; $\langle RFD \rangle_x^{x+50kb}$) with K562 MRT changes across 50 kb intervals ($\Delta_{50kb} MRT(x)$), following Eq. (2) with $v = 1.24 \text{ kb}\cdot\text{min}^{-1}$ assuming $T_S = 12$ hour. (B) Replication speed v derived from the proportionality coefficient (Eq. (2)) and (C) Pearson correlation coefficient (PCC) between the $\Delta_l MRT(x)$ and $\langle RFD \rangle_x^{x+l}$ profiles at the indicated scales l .

168 plateaued after ≈ 1 h (Arbona et al., 2018; Löb et al., 2016). Productive initiation events trap the
 169 factors at the forks and termination events release them, making them available for new initiation
 170 events. After grid search optimisation of the number of potential origins and the number of firing
 171 factors, we observed a high correlation (PCC = 0.88), similar to Gindin et al (Gindin et al. (2014);
 172 0.865), between simulated and experimental MRT profiles, and a lower correlation between sim-
 173 ulated and experimental RFD profiles (PCC = 0.70) (Fig. 3). Reasoning that addition of dispersed,
 174 random initiation to the PODLS (see Methods) might improve the results, we extended the grid
 175 search for this parameter and obtained an optimal correlation for RFD at PCC = 0.75 for 5% of ran-
 176 dom initiation events, while maintaining the same correlation with MRT (PCC = 0.88) (Fig. 3). These
 177 observations confirm that MRT and RFD data are consistent with each other and suggest that RFD
 178 data are furthermore informative about random initiation.

179 Despite the theoretical equivalence of MRT and RFD profiles (Eqs. (1) and (2)), their correlation
 180 (Fig. 2) decreased at small scales, due to the low (100 kb) resolution of MRT profiles. It also de-
 181 creased, to a lower extent, at large scales, because integrating RFD sums up its experimental noise.
 182 In fact, RFD provides better origin position information, while MRT better reflects integrated origin
 183 activity over broad regions. This is illustrated by the following numerical experiments.

184 When the positions of DNase I HSS were resampled within each 200 kb window prior to con-
 185 structing the PODLS, the simulated MRT profile retained a high correlation with the experimental
 186 MRT (PCC = 0.87; Fig. 4A, green curve), while the correlation between simulated and experimental
 187 RFD profiles dropped (PCC = 0.61; Fig. 4B, green curve). The exact positions of DNase I HSS were
 188 critical to reproduce RFD profiles upward jump positions, in line with the observed enrichment of
 189 OK-seq IZs with DNase I HSS (Petryk et al., 2016). On the other hand, the tolerance of MRT profiles
 190 to DNase I HSS resampling suggested that MRT is not sensitive to precise origin positions within a
 191 sampling window.

192 Although MRT can be computed by integrating RFD (Eq. (2)), this cumulates the experimen-

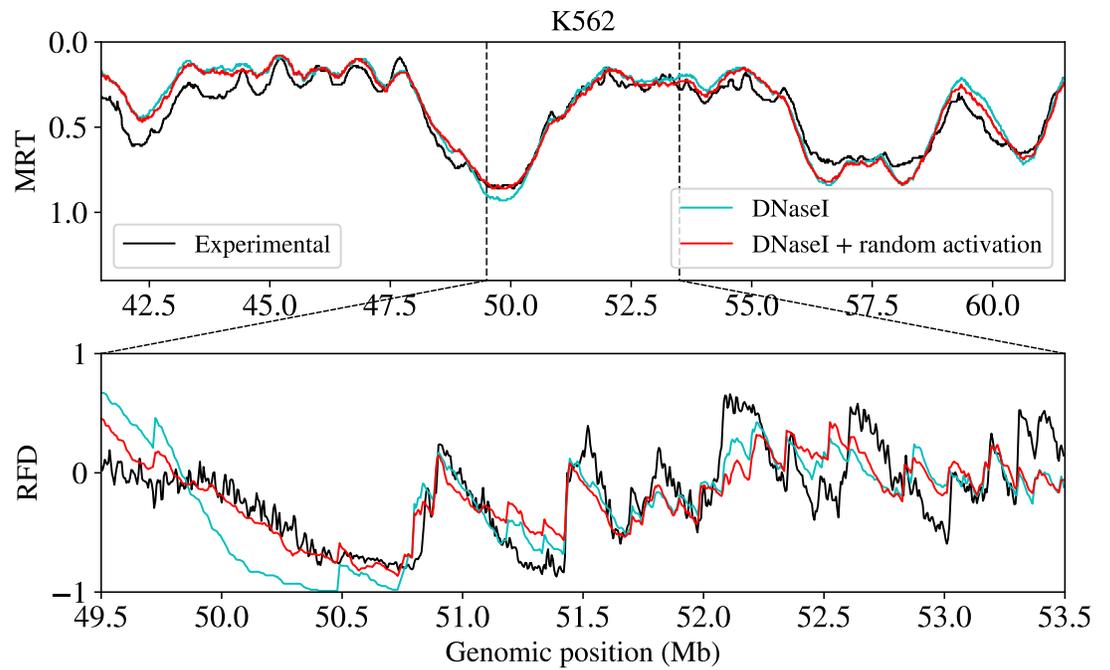


Figure 3. Comparison of experimental (black) and simulated (red, light blue) MRT profiles for a ≈ 20 Mb region of chromosome 1 (top) and RFD profiles for a 4 Mb region centered in the middle of the 20 Mb region (bottom) using a PODLS based on DNase I HSS, with (red) or without (light blue) the addition of 5% of random initiation events. All the parameters of the replication model except the percent of random initiation are the same.

193 tal noise, blurring large scale features that MRT data more directly capture. The lesser sensi-
 194 tivity of RFD than MRT to large scale patterns was revealed in a second numerical experiment
 195 where we modulated the DNase I HSS signal amplitude with a slow varying function of large pe-
 196 riod ($P = 25$ Mb) before constructing the PODLS. In that setting, the correlation between simulated
 197 and experimental profiles decreased markedly for MRT (PCC = 0.72) but only slightly for RFD (PCC
 198 = 0.72) (Fig. 4AC, blue curves). Therefore, MRT is constrained by the collective action of multiple
 199 origins, so that the sum of neighbouring DNase I HSS signals is critical, while their exact positions
 200 per 200 kb windows are not (Fig. 4A). RFD is instead sensitive to the location rather than the magni-
 201 tude of these signals. The influence of a single origin on RFD rapidly decreases with distance, due
 202 to the influence of intervening origins.

203 To summarize, incorporating RFD as a target for simulations likely allows to test origin position-
 204 ing at much higher resolution than was achieved with MRT (*Gindin et al., 2014; Löb et al., 2016*).
 205 Deriving RFD profile from MRT data (Eq. (1)) by numerical derivative would produce low resolution
 206 RFD profiles with amplified noise, while determining MRT profile from the summation of RFD data
 207 (Eq. (2)) would produce MRT profiles with unreliable MRT changes over large distances. Experimen-
 208 tal MRT and RFD profiles thus provide complementary information. We use both in the following
 209 analyses.

210 Learning a PODLS that accurately predicts both experimental MRT and RFD data.
 211 Having shown that experimental MRT and RFD profiles are consistent with each other over a wide
 212 range of scales at constant fork speed, we assessed to which extent they could be jointly explained
 213 by a single PODLS in our replication model. At 5 kb resolution, the PODLS correspond to ~ 575000
 214 parameters which must be optimised. To achieve this, we designed an iterative method that pro-
 215 gressively improves the PODLS (Fig. 5). It uses model simulations to train a neural network to pre-
 216 dict the PODLS given the MRT and RFD profiles, i.e., to invert our replication model. We initialised
 217 the method by setting non zero values of the PODLS in regions with the highest RFD derivative (See
 218 Materials and Methods) i.e. in the strongly ascending segments of the RFD profile corresponding

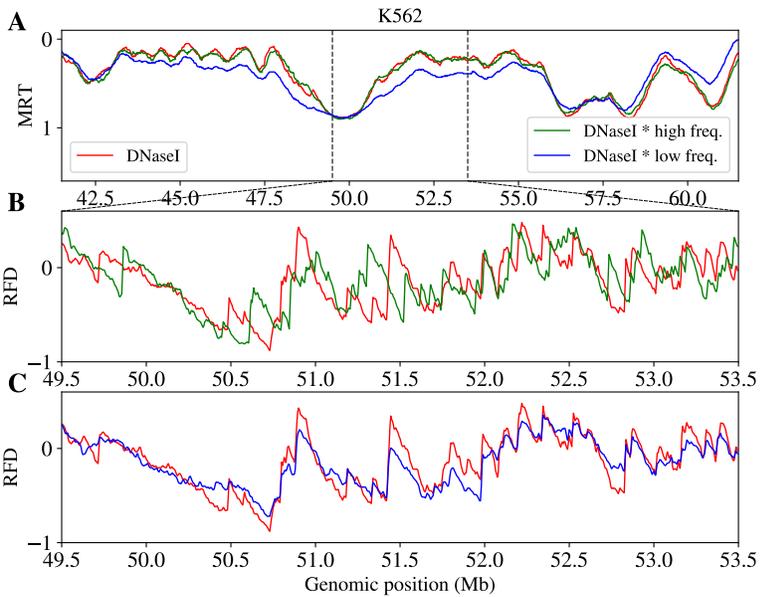


Figure 4. Comparison of simulated MRT (A) and RFD (B,C) profiles corresponding to different PODLS profiles all other model parameters being kept constant. PODLS were derived from (i) experimental DNase I HSS data (red), (ii) the DNase I HSS data after random shuffling of HS sites position within all 200 kb non-overlapping windows (green), and (iii) DNase I HSS data after modulating their amplitude over a period $P = 25$ Mb (we divided the amplitude signal by $1.1 + \cos(2x/P)$) (blue). DNase I HS site position shuffling in 200 kb windows does not influence simulated MRT profiles but alters RFD profiles significantly, as both red and green signals overlap in (A) but present clear differences in (B). Low-frequency modulation of HS site amplitude changes the relative strength of replication timing domains thus altering the MRT profiles, but do not influence the main features of the RFD profile as red and blue signal overlap in (C) but present clear differences in (A).

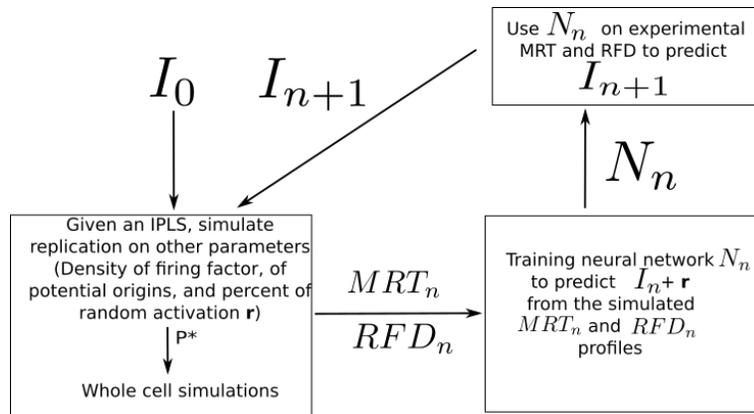


Figure 5. Schema of the iterative optimization procedure of the PODLS for simultaneous prediction of experimental MRT and RFD data. The starting PODLS I_0 may be a crude approximation of the target PODLS such as given by the peaks of RFD derivative, or the DNase HSS profile, but not a random profile. We observed that the procedure does not improve the prediction quality after a small number of iterations (maximum of 4 in *S. cerevisiae*, Supplementary Table S1).

219 to the main initiation zones previously described (Petryk et al., 2016). This crude approximation of
 220 the PODLS is named I_0 . Then a grid search optimisation on the other parameters $P = (\rho_F, d_{PO}, r)$
 221 of the replication model (See Material and methods) was performed. To limit computation time,
 222 this optimisation was performed over chromosome 2 only and resulted in a set of optimal paramet-
 223 ers P_0 that maximized the sum of the Pearson correlation coefficients between experimental and
 224 simulated MRT and RFD profiles. Then we simulated whole genome replication using I_0 and P_0 to
 225 generate MRT_0 and RFD_0 and trained a neural network (See Materials and Methods) to predict

226 $I_0 + r_0$ from MRT_0 and RFD_0 , where r_0 is the optimal fraction of random initiation events given I_0 .
227 We then used this network to predict I_1 from experimental MRT and RFD, reasoning that I_1 should
228 produce a more accurate prediction of MRT and RFD than I_0 . Another grid search optimisation
229 on P , given I_1 , was performed to obtain P_1 and given P_1 and I_1 we simulated MRT_1 and RFD_1 .
230 Then a new neural network was trained to predict $I_1 + r_1$ and was then applied to experimental
231 MRT and RFD to obtain I_2 . These steps were iterated four times, because the correlations between
232 experimental MRT and RFD profiles and their simulated estimates never improved with further
233 iterations.

234 We first applied the procedure to K562 data. The sequences of joint correlations between ex-
235 perimental MRT and RFD and simulated profiles (MRT_0, \dots, MRT_4) and (RFD_0, \dots, RFD_4) were (0.81,
236 0.93, 0.98, 0.98, 0.98) and (0.79, 0.89, 0.91, 0.92, 0.92), respectively. The highest joint correlation
237 was practically reached at the third iteration and we refer to the maximum initiation potential as
238 I_M . We ran a grid search on the whole genome given I_M , this yielded unchanged correlation for
239 both MRT and RFD, suggesting that parameter optimization on chromosome 2 only was not a lim-
240 itation. We also tried using K562 DNase I HSS as the I_0 of the method. The I_0 profiles obtained
241 from DNase I HSS or RFD derivative peaks presented some differences (PCC=0.76), but led to very
242 similar I_M profiles (PCC=0.94; Supplementary Figure S2) and produced identical high correlations
243 between simulated and experimental MRT (0.98) and RFD (0.91) profiles. In contrast, we were un-
244 able to ameliorate the PODLS starting from a random I_0 (MRT and RFD correlations were 0.67 and
245 0.84, respectively, using I_2 , but decreased at step 3). Therefore, our optimization method required
246 some initial information about the PODLS, but converged to nearly the same optimized PODLS
247 from heterogeneous starting points. This is not a constraint as an adequate initialisation can be
248 obtained from experimental RFD data.

249 To test the robustness of this inversion procedure, it was applied to replication profiles of
250 GM06990 and HeLa human cell lines and yeast *Saccharomyces cerevisiae*. It systematically resulted
251 in high PCC between experimental and simulated profiles at the third or fourth iteration: for MRT
252 0.99 with GM06990; 0.99 with HeLa and 0.96 with *S. cerevisiae*; for RFD 0.91 with GM ; 0.84 with
253 HeLa RFD and 0.91 with *S. cerevisiae* (see Supplementary Table S1 for the results of the different
254 iterations). Figure 6 illustrates the striking consistency between simulation and experiments ob-
255 tained for the three different human cell lines. Correlation for HeLa RFD profile was less than for
256 other cell lines. Indeed HeLa is more challenging as it has about twice as many IZs as K562 and
257 GM06990 (Petryk et al., 2016; Wu et al., 2018), but regions of poor RFD prediction also showed
258 inconsistencies between experimental MRT and RFD probably due to the use of different HeLa cell
259 lines in different laboratories (Fig. 6).

260 Sensitivity of MRT and RFD with respect to model parameters.

261 For K562, we performed simulations of whole genome replication to assess the sensitivity of MRT
262 and RFD and S-phase duration to: (i) the density of firing factors ρ_F (number per Mb); (ii) the mean
263 distance between potential origins d_{PO} ; (iii) the proportion of random initiation r ; and (iv) fork speed
264 v . Working around the reference set of parameters ($\rho_F = \rho_F^* = 0.56 \text{ Mb}^{-1}$; $r = r^* = 0\%$, $d_{PO} = 20$
265 kb, $v = 1.5 \text{ kb} \cdot \text{min}^{-1}$), we let one of the parameter vary (Fig. 7). ρ_F^* and r^* are the optimal values
266 obtained at the end of the iterative procedure, without the final grid search exploration on the
267 whole genome as it did not improve the correlations; the values for d_{PO} and v are reasonable
268 choices justified below.

269 The correlations of simulated with experimental MRT and RFD profiles showed a clear maxi-
270 mum at $\rho_F = \rho_F^* = 0.56 \text{ Mb}^{-1}$, more pronounced for RFD (Fig. 7A,B). This number implied that the
271 maximum density of forks was ~ 1 fork per Mb, or $\sim 6,000$ forks per diploid nucleus, at any time
272 in S phase. This was in reasonable agreement with upper estimates of the maximal density of
273 limiting factor Cdc45 (3.5 molecules per Mb) in mammalian cells (Wong et al., 2011), considering
274 that 2 molecules of Cdc45, but only one limiting factor in our model, are required to activate two
275 diverging forks.

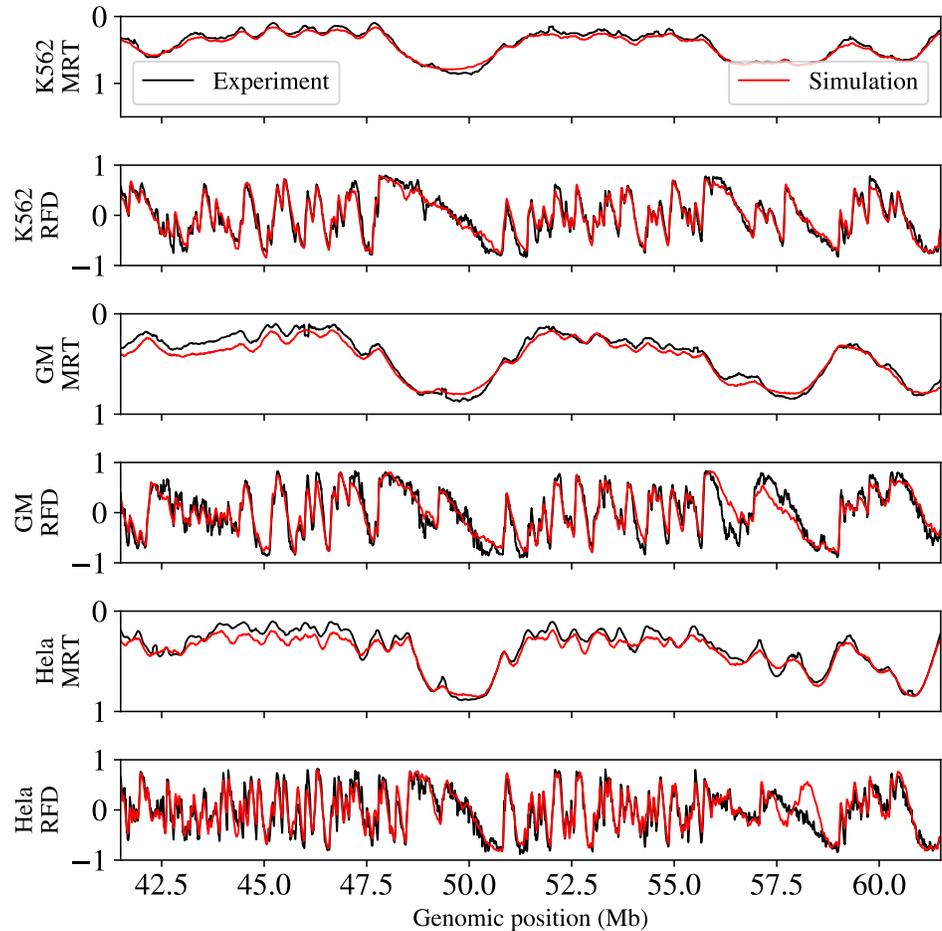


Figure 6. Comparison for the three indicated cell lines of experimental MRT and RFD (black) and simulated MRT_M and RFD_M profiles (red). A representative 20 Mb segment of chromosome 1 is shown. Note that the HeLa region of strong discrepancy between experimental and simulated profiles (between 57.7 and 58 Mb) correspond to a region where experimental MRT and RFD are not coherent: there is no increasing RFD region corresponding to the early timing peak at position 57.6 Mb.

276 The optimal value for the random initiation $r^* = 0\%$ was confirmed as increasing r up to 20%
 277 slightly decreased both MRT and RFD correlations (Fig. 7G,H). The value $r = 0\%$ means that random
 278 initiation was correctly learned by the iterative method and did not require further external addition.
 279 In order to further apprehend the requirement of a minimal amount of random initiation,
 280 we performed the following experiments: (i) the random initiation was treated as an external parameter
 281 in the iterative PODLS optimisation procedure i.e., it was excluded from the training, the
 282 neural network was trained to output I_n instead of $I_n + r$; in that case the correlation with MRT
 283 and RFD was maximal for 5% of added random initiation (Fig. 7G,H); (ii) setting to zero the lowest
 284 I_M bin values totalling 5% of potential origins ($\approx 53\%$ of the bins), significantly decreased the correlation
 285 with RFD data to $PCC = 0.82$: without these random initiations, the simulations failed to
 286 capture the RFD in large late replicating regions whose replication time was also clearly delayed
 287 (Supplementary Figure S5). The extended null RFD segments observed in these regions are indeed
 288 consistent with random initiation.

289 Varying d_{pO} and v over a large range of values only weakly changed the correlation of experimental
 290 and simulated MRT and RFD (Fig. 7D,E,J,K), which justifies why they were left out of the
 291 parameter optimization procedure. We decided to set v to $1.5 \text{ kb}\cdot\text{min}^{-1}$ and d_{pO} to 20 kb, for the

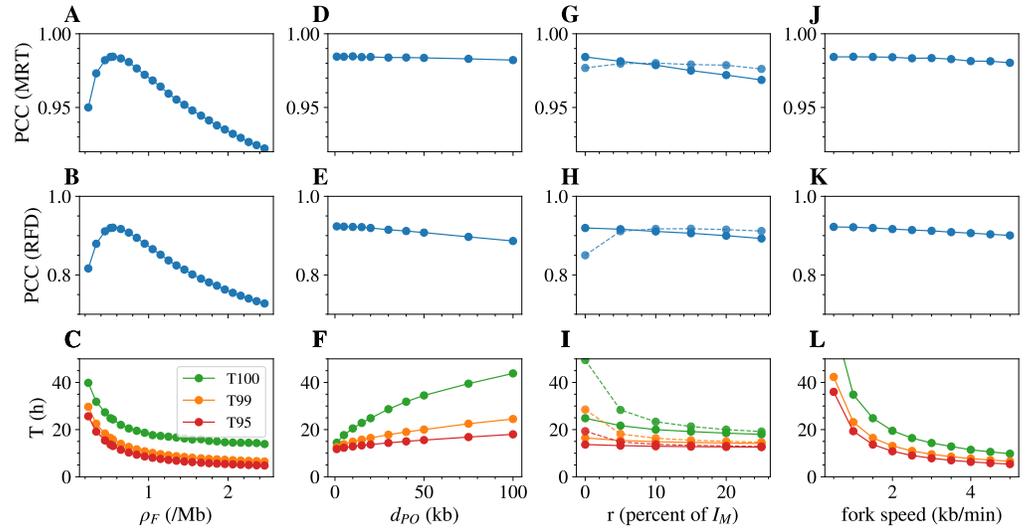


Figure 7. Effect of single parameter variation on measurable targets in K562, other parameters being kept as their value in the reference parameter set. Effect of the density of firing factors ρ_F (A,B,C); the average distance between potential origins d_{PO} (D,E,F); the proportion of random initiation r (G,H,I); the fork speed v (J,K,L), on the Pearson Correlation Coefficient (PCC) between simulated and experimental MRT (A,D,G,J) and RFD (B,E,H,K) profiles, and on the median of T95 (red), T99 (orange) and T100 (green), the times required to replicate 95% 99% and 100% of the genome (C,F,I,L). In (G,H,I) dots joined by dashed lines correspond to the effect of the additional random activation r when using the PODLS profile determined using a neural network where r is treated as an outside parameter (the network is trained on I_n and not on $I_n + r$), see main text.

292 following reasons. First, single molecule studies of DNA replication in human cells have repeat-
 293 edly reported replication fork speeds of 1-3 kb min⁻¹ and distances between activated origins of
 294 50-200 kb (Conti *et al.*, 2007; Técher *et al.*, 2013). Second, MCM depletion experiments indicated
 295 a 5-10 fold excess of loaded MCM DHs over actual initiation events (Ibarra *et al.*, 2008). Third,
 296 biochemical quantitation suggested that chromatin is loaded with 1 MCM DH per 20-40 kb at S
 297 phase entry (Burkhart *et al.*, 1995; Wong *et al.*, 2011). Taken together, these figures are reason-
 298 ably consistent with each other and with a d_{PO} of 20 kb. Similar results were robustly observed
 299 using GM06990 and HeLa replication data, with maximum PCC values observed for ρ_F^* of 0.56 and
 300 0.91 Mb⁻¹, respectively (Supplementary Figures S6 and S7).

301 In summary, MRT and RFD data, being highly consistent with each other, are jointly and pre-
 302 cisely explained by a simple model featuring a unique PODLS input and values for d_{PO} , v and ρ_F in
 303 agreement with the current knowledge. A small amount $\sim 5\%$ of random initiation was necessary
 304 to fully account for experimental data, suggesting that most if not all of the genome has a minimal
 305 replication initiation potential.

306 Dependence of replication kinetics on model parameters.

307 We first analyzed K562 S-phase duration T_S using the median of the times to replicate 95%, 99%
 308 and 100% of the genome, T95, T99 and T100 respectively. As expected, each of these three times
 309 decreased with the density of firing factors ρ_F and the fork speed v (Fig. 7C,L) as we are in a regime
 310 of strong affinity between firing factors and potential origins (large k_{on}) so that $T_S \approx \frac{1}{2*v*\rho_F}$ (Arbona
 311 *et al.*, 2018). The model predicted much larger differences between T100 and T99, than between
 312 T99 and T95, consistent with the latest replicated regions being the most devoid of potential origins.
 313 Indeed, for a genome-averaged distance $d_{PO} = 20$ kb, the predicted distance between potential
 314 origins increased from a short 2 kb value in MRT < 0.15 regions to 380 kb in MRT > 0.85 regions
 315 (Supplementary Figure S3). This observation also explained that (i) the cell-to-cell variability of
 316 T95 or T99 (~ 10 min) was much smaller than that of T100 (hours) (Supplementary Figure S4); (ii)
 317 increasing d_{PO} increased T100 to a much greater extent than T95 or T99 (Fig. 7F); and (iii) adding

318 random initiation decreased T100 to a much greater extent than T99 or T95 (Fig. 7I). The latter effect
319 was maximal when random initiation was an outside parameter, and decreased with increasing
320 r , consistent with the latest replicated regions being fully devoid of origins when $r = 0$ (Fig. 7I).
321 Consistently, many late replicating regions show flat MRT and null RFD profiles revealing random
322 initiation (Petryk et al., 2016), but the even later-replicating, common chromosomal fragile sites
323 (CFSs), show an origin paucity that explains their fragility (Letessier et al., 2011).

324 Experimentally reported S phase lengths were closer to T95 than T100. Using the reference set
325 of parameters ($\rho_F = \rho_F^*$; $r = r^* = 0\%$, $d_{PO} = 20$ kb, $v = 1.5$ kb.min⁻¹), the predicted T95 was 8.6 h for
326 HeLa (experimental estimate 8.8 h; Hahn et al. (2009)), 13 h in K562 (experimental estimate 12 h;
327 Weis (2012)) and, taking account a fork speed of $v = 2.0$ kb.min⁻¹ in the closely related JEFF cell
328 line (Técher et al., 2013), 10.7 h for GM06990 (experimental estimate 10 h; Guilbaud et al. (2011)).
329 One probable explanation is that experimental detection of S phase markers misses the earliest
330 and latest S phase cells, when the replication rate is the lowest. Indeed, very late replication of
331 specific sequences was reported to linger during what is termed the G2 phase of the cell cycle
332 (Widrow et al., 1998), and S phase length variations around the mean ranged from minutes to
333 hours depending on cell lines and detection methods (Pereira et al., 2017; Weber et al., 2014).
334 Within the parameter range we explored, T95 variations (~10 min) were smaller than for T100
335 (hours) (Fig. S4). Experiments may therefore have underestimated the exact duration of S phase.
336 Another possibility is that we underestimated r or v , since increasing either parameter above its
337 reference value efficiently reduced T100 without much compromising the correlations of simulated
338 and experimental MRT and RFD (Fig. 7GHI, Fig. S6, Fig. S7). In our simulations, the faster replication
339 of HeLa cells was explained by a larger number of firing factors than in GM06990 (2.0-fold) and K562
340 (1.9-fold). Thus, the optimisation procedure selected a density of firing factors ρ_F that gave relative
341 S phase durations consistent with experimental measurements using sensible values for d_{PO} and
342 v .

343 The origin firing rate per length of unreplicated DNA, $I(t)$, was previously reported to follow a
344 universal bell-shaped curve in eukaryotes (Goldar et al., 2009; Arbona et al., 2018). As expected
345 from the choice of the k_{on} value, the simulations produced the expected shape for the three cell
346 lines with a maximum of $I(t)$, I_{max} between 0.01 and 0.02 Mb⁻¹.min⁻¹, in reasonable agreement with
347 a crude estimate of 0.03-0.3 Mb⁻¹.min⁻¹ from low-resolution MRT data (Supplementary Figure S8)
348 (Goldar et al., 2009). Finally, we measured in K562 the dispersion of replication times (RT) of each
349 locus as a function of its MRT. A recent high-resolution Repli-Seq study (Zhao et al., 2020) reported
350 that RT variability, estimated as the width between the first and the third quartiles of RT distribution,
351 increased from early to mid S phase and decreased thereafter. For most regions the RT variability
352 was in the 1.25–2.5 h range. Our simulations in K562 produced a similar behavior of RT variability
353 but over a wider range, from ~ 0.5 h in early or late S phase to 3 h in mid S phase (Supplementary
354 Figure S10).

355 In summary, the kinetic parameters of S phase predicted by our stochastic model were (i) con-
356 sistent with the reported time-dependencies of the firing rate $I(t)$ and the RT variability and (ii)
357 predictive of relative S phase durations. However, RT variability was broader than reported, sug-
358 gesting lower stochasticity of *in vivo* replication kinetics than in our model, in which origins fire
359 strictly independently of each other.

360 **Direct estimation of the PODLS from experimental data**

361 Origin firing can be prevented by context-dependent passivation from nearby origins. An impor-
362 tant concept in DNA replication modeling is therefore the distinction between observed origin effi-
363 ciency (OE) and intrinsic origin efficiency (IE). OE is the fraction of origin copies that fire in a popu-
364 lation, whereas IE is the efficiency that would be observed, in the absence of passivation, over the
365 entire length of S phase.

366 In our model, potential origins are MCM DHs which all have the same elementary probability
367 of firing. For one bin x with $n(x)$ MCM DHs, given the reaction rate k_{on} and the number of free firing

368 factor $F_{free}(t)$, the probability for firing to take place during an elementary time dt is:

$$k_{on}n(x)F_{free}(t)dt . \quad (3)$$

369 If we consider that $F_{free}(t)$ is constant and equal to $[F_{free}]$ for a large part of the S phase (Supplemen-
370 tary Figure S11), the probability $A_x(t)$ for bin x to have been activated at time t without considering
371 passivation is given by:

$$A_x(t) = 1 - e^{-k_{on}n(x)[F_{free}]t} . \quad (4)$$

372 Hence, with an infinite time to fire, $A_x(t)$ converges to one unless the locus is devoid of any potential
373 origins ($n(x) = 0$). The observed firing efficiency OE is smaller than one due to passivation. Reason-
374 ing that replication of a small bin occurs much more often by passivation from nearby origins than
375 by internal initiation (for 5 kb bins, $\Delta RFD/2$ which is an approximation of OE as discussed later
376 has a maximum value of 0.17), we can consider that the average passivation time of a bin is not
377 very different from its MRT and thus $OE(x) = A_x(MRT(x))$ which leads to:

$$OE(x) = 1 - e^{-k_{on}n(x)[F_{free}]MRT(x)} . \quad (5)$$

378 We first assessed the validity of this relationship in the set of 200 S-phase simulations using the
379 optimized PODLS I_M and the reference set of model parameters in K562. $OE(x)$ and $MRT(x)$ were
380 the averages of origin firing status and RT recorded in each simulation. The total number of MCM
381 DHs over the genome being L/d_{PO} and as the I_M profile is normalised to 1, the average MCM DH
382 density profile is $n(x) = I_M(x) * L/d_{PO}$ (MCM DH/5kb). Finally we used $[F_{free}] = 52$, the almost
383 constant value observed for $MRT < 0.5$ in K562 simulations (Supplementary Figure S11). The
384 two terms of Eq. (5) were computed. They showed a genome-wide Pearson correlation coefficient
385 of 0.83 and a proportionality coefficient of 0.96. When focusing on the local maxima of ΔRFD
386 ($\sim 10,000$ peaks in K562, Materials and methods), which correspond to IZs, the PCC raised to 0.9,
387 and the proportionality coefficient to 0.98 (Supplementary Figure S12). These results indicate that
388 our hypothesis that $OE(x) = A_x(MRT(x))$ is globally valid but applies even more precisely at IZs
389 determined as ΔRFD local maxima. However, when exploring different values for k_{on} and d_{PO} , we
390 noted that the PCC was stable, but the proportionality coefficient varied from 0.6 to 1.7, even if we
391 replaced the assumed constant $[F_{free}]$ by its time-dependent value $F_{free}(t)$. This indicates that, for
392 unclear reasons, the validity of Eq (5) is sensitive to the precise combination of origin density and
393 reactivity parameters.

394 Generally speaking, the ΔRFD across a genomic segment is twice the difference between the
395 density of initiation and termination events in the segment (Audit et al., 2012). Mammalian IZs are
396 broad and may contain multiple MCM DHs. Termination events are nevertheless rare or absent
397 within IZs (Petryk et al., 2016; Hamlin et al., 2010; Blin et al., 2021). Therefore, forks emanating
398 from the first activated MCM DH must rapidly passivate nearby MCM DHs and one can estimate
399 $OE(x) = \Delta RFD(x)/2$ in these loci. For example, the RFD will shift from -1 to +1 across a 100%
400 efficient IZ, creating a jump of $\Delta RFD = 2$. We indeed found a PCC of 0.96 with a proportionality
401 coefficient of 0.92 between $OE(x)$ and $\Delta RFD(x)/2$ at the 10,000 local maxima of ΔRFD in our
402 K562 simulations (Fig S13). As a consequence, Eq. (5) linking $OE(x)$, $n(x)$ and $MRT(x)$ provides a
403 link between $RFD(x)$, $n(x)$ and $MRT(x)$ in IZs suggesting a simple and direct way to estimate the
404 PODLS ($\propto n(x)$) from MRT and RFD profiles:

$$n_e(x) = \frac{-\ln(1 - \Delta RFD(x)/2)}{k_{on} [F_{free}] MRT(x)} , \quad (6)$$

405 where $n_e(x)$ is the $n(x)$ profile estimated from the measurable parameters ΔRFD and MRT. Note
406 that for small value $\Delta RFD(x)$ (for 5 kb bins, the maximum value of $\Delta RFD/2$ is 0.17), one can use a
407 Taylor extension of \ln to simplify Eq. (6) yielding:

$$n_e(x) \approx \frac{\Delta RFD(x)}{2k_{on} [F_{free}] MRT(x)} . \quad (7)$$

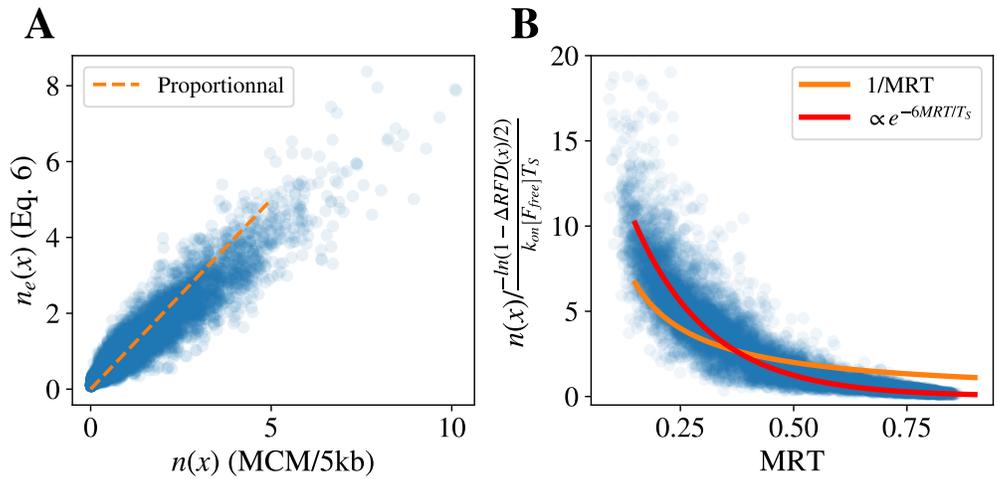


Figure 8. (A) Comparison of the average number $n(x)$ of MCM per 5kb bin used in the simulations, and the predicted number of MCM $n_e(x)$ using Eq. (6) at the 10,000 ΔRFD peaks in K562. (B) Scaled $n(x) \times k_{on}[F_{free}]T_s / (-\ln(1 - \Delta RFD(x)/2))$ as a function of Repli-seq MRT at the 11,000 ΔRFD peaks (blue circles) in K562 simulations; it is compared with $1/MRT$ (orange) and proportionality to \exp^{-6MRT/T_s} (red).

408 We then compared $n_e(x)$ estimated using simulated MRT and RFD with $n(x)$, the input of the simu-
 409 lation (Fig. 8 A). The Pearson correlation was 0.95 and the coefficient of proportionality was 0.78
 410 at the 10,000 ΔRFD local maxima. Similarly to Eq. (5), when exploring different values for k_{on} and
 411 d_{PO} , the PCC remained stable but the proportionality coefficient varied from 0.4 to 1.2.

412 We observed from further empirical exploration that the dependency of $n(x)$ on $MRT(x)$ (Fig. 8
 413 B) was even better captured (PCC = 0.97) by the exponential dependency on MRT of Eq. (8) than by
 414 the inverse linear dependency of Eqs. (6) :

$$n_e^{exp}(x) \propto \Delta RFD(x)e^{-6MRT(x)/T_s} \quad (8)$$

415 We hypothesise that Eq. (8) accounts for the actual variations of $F_{free}(t)$ (Supplementary Figure S11)
 416 boosting potential origin firing efficiency in late S phase, so that a smaller number of MCM DHs are
 417 required to produce the same OEs.

418 We then wondered if it would be possible to simulate replication using a PODLS directly de-
 419 rived from experimental data using Eqs. (6) or (8) (note that since the PODLS is normalized the
 420 ignorance on the prefactors in Eq. (8) is not an issue). For this we selected the 15% highest exper-
 421 imental $\Delta RFD(x)$ (Materials and Methods) and used Eq. (6) to predict the PODLS that we used in
 422 our model. The resulting simulated MRT and RFD profiles were highly correlated with experimen-
 423 tal profiles (Table 1), for example PCC=0.94 between MRT and 0.88 between RFD profiles in K562.
 424 Interestingly, this formula for PODLS prediction robustly applied also in *S. cerevisiae* with Pearson
 425 correlations of 0.93 for MRT and 0.90 for RFD. The PODLS derived from Eq. (8) led to even higher
 426 correlations between simulated and experimental MRT (PCC=0.97) and RFD (PCC=0.91), very close
 427 to the correlation coefficients obtained using I_M profiles (Table 1). We also confirmed this result in
 428 *S. cerevisiae* with PCC=0.96 for MRT and PCC=0.9 for RFD. These results show that combining OEs,
 429 estimated by local RFD upshifts, with MRT data, suffices to produce a near optimal PODLS.

430 SNS-seq or Bubble-seq signals also provide in principle a direct estimate of OEs, up to a pro-
 431 portionality coefficient. Using such OE estimates (K562 and HeLa SNS-seq, GM 06998 Bubble-seq)
 432 directly as PODLS resulted in poor correlations between simulated and experimental profiles (MRT,
 433 PCC=0.54 , 0.43 and 0.30 and RFD, PCC=0.16 , 0.23 and 0.12, respectively). Combining the same
 434 data with MRT data to infer the PODLS using Eq. (8) improved the correlations (MRT, 0.81 , 0.87
 435 and 0.83 and RFD 0.48 , 0.59 and 0.53, respectively) (Table 1 1), but combining the same MRT infor-
 436 mation with a flat OE profile produced even better correlations (MRT, 0.97 , 0.98 and 0.98 and RFD

437 0.71 , 0.62 and 0.71, respectively), suggesting that SNS and Bubble-seq data do not synergize with
438 MRT as favorably as RFD data.

439 We also analysed experimental data on potential origin positioning (Table 1). Using PODLS
440 computed from K562 ORC2 (*Miotto et al., 2016*), HeLa MCM7 (*Sugimoto et al., 2018*) and HeLa
441 MCM2 (*Foss et al., 2021*) resulted in PCCs between simulated and experimental profiles of 0.87,
442 0.46 and 0.54 for MRT and 0.74, 0.28 and 0.43 for RFD, respectively. Finally, we computed the
443 correlation of SNS-seq, Bubble-seq and ORC/MCM signals with the optimized PODLS I_M in the
444 corresponding cell line at 5 kb and 50 kb resolution. For K562 SNS, HeLa SNS and GM Bubbles, the
445 correlations at 5 kb and 50 kb were 0.05 and 0.16, 0.19 and 0.32, and 0.06 and 0.15, respectively. For
446 K562 ORC2, HeLa MCM7 and HeLa MCM2, the correlations were 0.30 and 0.54, 0.19 and 0.32, and
447 0.27 and 0.42, at 5 kb and 50 kb resolution, respectively. Therefore, in contrast to RFD upshifts,
448 none of these experimental datasets were convincing predictors of the PODLS I_M , for reasons
449 that remain to be elucidated. The fact that ORC location better predicted the PODLS than MCM
450 is unexpected if the multiple MCM DHs loaded by ORC are equally competent to trigger initiation
451 (*Harvey and Newport, 2003; Edwards et al., 2002*). Assuming this discrepancy does not stem from
452 experimental limitations, it suggests that ORC-proximal MCM DHs are more likely to fire than ORC-
453 distal ones.

454 In our model, all origins have the same k_{on} , and the spatial dependency is encoded in the non
455 uniform density of potential origins. Since the effective reactivity of a potential origin is propor-
456 tional to $k_{on}F_{free}$, the observed differences between the experimental MCM density and the inferred
457 PODLS (I_M) may be explained by spatial or temporal non-uniformity, i.e. locus-dependent k_{on} or
458 time-dependent F_{free} , with the $k_{on}F_{free}$ landscape given by the I_M/MCM ratio. We found that the
459 normalized I_M/MCM ratio computed in 50 kb windows in HeLa cells (excluding null MCM win-
460 dows; Fig. S14) decreased with MRT but was broadly dispersed even at constant MRT. The global
461 trend could be explained if firing factor abundance decreased during S phase, as recently reported
462 (*Wittig et al., 2021*), but the broad dispersion also implied that even at similar MRT, all MCM DHs
463 are not equally reactive to firing factors, possibly due to MCM DH modifications or to chromatin
464 environment. Finally, we cannot exclude that experimental noise or differential MCM loading dy-
465 namics during G1 (*Mei et al., 2021*) prevent an accurate picture of MCM distribution at S phase
466 entry. Interestingly, the I_M/ORC ratio (computed in K562) did not vary with MRT but was still
467 broadly dispersed at constant MRT Fig. S14). This suggests that more MCM DHs are loaded per
468 ORC in late than in early replicating regions, but that the resulting equalization of MCM loading is
469 counteracted by above-discussed mechanisms, possibly including an increased firing propensity
470 of ORC-proximal MCM DHs.

471 According to Eq. (6), the faster passivation of early than late IZs means that early IZs require
472 several-fold more MCM DHs than late IZs to achieve a similar OE. In our simulation with optimized
473 k_{on} , F_{free} and d_{PO} , the maximum RFD upshift per 5kb bin was 0.17, an OE that would require as much
474 as ~ 20 MCM DHs per 5 kb if an early MRT of 1h is to be achieved. Given that MCM DHs occupy 60
475 bp each and are only found in internucleosome linkers *Foss et al. (2021)*, a 5 kb chromatin segment
476 may not accommodate more than 25 MCM DHs. This steric limit is almost reached (see also Fig S3),
477 suggesting that additional regulatory mechanisms that increase k_{on} or F_{free} in early S phase may
478 be required to boost the intrinsic firing efficiency of some MCM DHs, consistent with Fig. S14.

479 To summarize, we checked that OEs can be accurately measured from RFD upshifts and we
480 could predict from these OEs and MRT the number of potential origins (MCM DHs) per 5 kb bin
481 in IZs assuming the specific value for d_{PO} used in our model, where all MCM DHs have the same
482 locus- and time-independent probability of firing per unit time. However, the observed discrepan-
483 cies between predicted and observed MCM DH densities, and the steric MCM loading constraint
484 discussed above, support a mixed, potential origin density and affinity model where MCM DHs
485 may have different affinities for firing factors. Future investigations of the I_M/MCM and I_M/ORC
486 ratios should help reveal the licensing and post-licensing mechanisms that regulate origin firing
487 probability.

PODLS	K562 MRT	K562 RFD	GM MRT	GM RFD	Hela MRT	Hela RFD
I_M	0.98	0.92	0.99	0.91	0.99	0.84
n_e^{exp} Eq. (8)	0.97	0.91	0.97	0.88	0.96	0.83
n_e Eq. (6)	0.94	0.88	0.91	0.82	0.92	0.80
e^{-6MRT/T_S}	0.97	0.71	0.98	0.71	0.98	0.62
ORC2	0.87	0.75	-	-	-	-
SNS e^{-6MRT/T_S}	0.81	0.48	-	-	0.87	0.59
Bubble e^{-6MRT/T_S}	-	-	0.83	0.53	-	-
MCM2	-	-	-	-	0.54	0.43
MCM7	-	-	-	-	0.46	0.28
SNS	0.54	0.16	-	-	0.43	0.23
Bubble	-	-	0.30	0.12	-	-

Table 1. Best joint correlation between simulated and experimental MRT and RFD data in K562, GM and Hela cell lines, marginalising over the other parameters of the simulation for different choices of PODLS.

488 Discussion

489 Without noise, MRT and RFD profiles in a constant fork speed hypothesis account for the same
490 information. Here we have shown that it is possible to compare MRT increments with the integral
491 of RFD with only one free parameter, the fork speed. The narrow range of values obtained by
492 fitting the fork speed (from ≈ 1.2 to ≈ 1.6 kb/min) over the large range of scales explored (5 kb
493 to 5 Mb), and the high correlations obtained (from 0.68 to 0.95), suggest that both experiments
494 are compatible, even at lower resolutions than expected (MRT resolution ≈ 100 kb) and that the
495 hypothesis of a constant fork speed at resolutions down to 5 kb is robust. We have also shown by
496 randomly resampling DNaseI HSSs that RFD profiles contain higher resolution information than
497 MRT profiles. However MRT profiles contain information about integrated initiation strengths of
498 large domains, that is lost when integrating noisy RFD profiles. We therefore used both profiles in
499 our inversion method.

500 Mathematical models have been developed to estimate intrinsic origin efficiencies from MRT
501 (*de Moura et al., 2010; Baker and Bechhoefer, 2014*), and RFD (*Bazarova et al., 2019*). The models
502 assign either a discrete number of origins (*de Moura et al., 2010; Bazarova et al., 2019*), each having
503 a time-dependent probability of firing, or a continuous spatiotemporal initiation density (*Baker and*
504 *Bechhoefer, 2014*). In principle our inversion method could be applied to these models, but their
505 hypotheses are more complex than ours. As the probability to activate an origin changes with
506 time, it is more difficult to test whether firing efficiency is set at the end of licensing. Furthermore,
507 these methods either require a non trivial optimisation that may be feasible with the yeast but
508 not the human genome, given its size, or a Bayesian analysis that so far was limited to sets of 3
509 origins *Bazarova et al. (2019)*. In contrast, our approach is very flexible, can easily accommodate
510 new datasets and is fast even with the human genome. Furthermore it outputs a 1D profile of
511 potential origin density in human cells, which can be directly compared to experimental origin
512 licensing profiles, as first achieved in yeast (*Das et al., 2015*).

513 We found that the PODLS could be segmented in peaks and flat areas of random initiation.
514 However the level of random activation found here (5 %) was much lower than inferred in Miotto
515 (*Miotto et al., 2016*) (60%). Our estimate is more consistent with recent single molecule analyses
516 reporting 10-20% of random initiation events in yeast (*Müller et al., 2019; Hennion et al., 2020*).
517 Nevertheless, the correlations between simulated and experimental MRT and RFD data were not
518 much affected by increasing the percentage of initiation and we cannot exclude that our estimate
519 is conservative. The inferred PODLS allowed us to generate simulated MRT and RFD profiles ex-

520 tremely similar to experimental ones in three different cell lines. The correlations were 0.98 for
 521 MRT and 0.86-0.91 for RFD. Therefore, a remarkably simple stochastic model featuring a constant
 522 fork speed, an almost constant number of limiting firing factors and a time-independent initiation
 523 strength profile suffices to jointly account for MRT and RFD data nearly exactly. Importantly, the
 524 correlations obtained using experimental ORC or MCM-based PODLSs were lower than with the
 525 optimal PODLS I_M inferred by neural networks, by 0.11-0.16 for ORC and by 0.4-0.5 for MCM. This
 526 suggests either that ORC and, more strikingly, MCM datasets do not accurately reflect the true
 527 distribution of these proteins at S phase entry, or that their abundance is not the sole factor deter-
 528 mining initiation strength.

529 Reasoning that an origin's passivation typically occurs at its MRT, we derived a novel mathe-
 530 matical relationship (Eq. (5)) that, assuming a constant availability of firing factors through S phase,
 531 predicts an origin's intrinsic efficiency (IE) from its observed efficiency (OE) and its MRT. OEs can
 532 be estimated by RFD upshifts. Knowing d_{PO} , predicted IEs can be converted into absolute MCM DH
 533 densities proportional to the RFD upshift and to $1/MRT$ (Eq. (6)). This was fairly well verified in
 534 simulated datasets, but we empirically found that an $1/e^{6MRT}$ dependency (Eq. (8)) gave even bet-
 535 ter predictions, probably because the recycling of firing factors by termination events increases
 536 in late S phase. Strikingly, the PODLS inferred from experimental RFD and MRT data using Eq. (8)
 537 generated almost as good simulated profiles as the PODLS inferred using neural networks. We
 538 however caution that this procedure is sensitive to smoothing and thresholding parameters and
 539 is therefore less robust than neural networks.

540 Eq. (5) to (8) imply that early IZs require several-fold more MCM DHs than late IZs to achieve a
 541 similar OE, which may potentially explain why SNS-seq and Bubble-seq profiles were poorly corre-
 542 lated to the optimal PODLS. We therefore used Eq. (8) to exploit the MRT information and convert
 543 these OE measurements into IEs. The resulting PODLS was improved but gave poorer results than
 544 a PODLS inferred from a flat OE profile. We conclude that SNS-seq and Bubble-seq data are not
 545 accurately consistent with MRT and RFD data.

546 Although our model allowed us to infer a PODLS that jointly predicts RFD and MRT almost ex-
 547 actly, this PODLS was not perfectly correlated with experimental ORC and MCM profiles. Assuming
 548 that these profiles are not biased, this imposes to relax the assumption that all MCM DHs are
 549 equally reactive to firing factors, or the assumption that the concentration of firing factors is con-
 550 stant. Examination of the I_M/ORC and I_M/MCM ratios suggests that, while more MCMs per ORC
 551 are loaded in late- than in early-replicating DNA, MCM DH activation probability is higher in early-
 552 replicating DNA and next to ORC, and varies along the genome even at constant MRT. Mechanisms
 553 that increase I_M/MCM may allow early IZs to reach a high IE without increasing MCM DH density
 554 beyond steric constraints. Understanding the genetic and epigenetic determinants of MCM DH
 555 density and reactivity to firing factors is a goal for future studies.

556 Materials and Methods

557 Model and simulation

558 Model and parameters

559 We model the replication initiation process by a bimolecular reaction between free firing factors
 560 F_{Free} and potential origins PO on the genome of size L with a reaction rate k_{on} . This rate is the
 561 probability per unit of time that a firing factor and a PO meet and that an initiation follows. Once
 562 an initiation event has occurred, two forks propagate in opposite direction at speed v and a firing
 563 factor is trapped. The number N_F of firing factor is fixed and parametrised as $N_F = \rho_F L$ with ρ_F the
 564 density of firing factors. A potential origin density landscape I is used to position $\frac{L}{d_{PO}}$ origin prior
 565 to each S phase entry along the genome, with d_{PO} the genome average distance between potential
 566 origins. We also decompose the initiation probability landscape I in two terms: it is the sum of
 567 an inhomogeneous profile I_e , for example derived from an epigenetic landscape, and a uniform
 568 contribution which correspond to a proportion r of the initial profile I_e . In all the simulations we

569 fixed $v = 1.5$ kb/min and $k_{on} = 3e^{-6}$ min⁻¹. A typical simulation therefore has 4 free parameters
570 ($\rho_F, \rho_{PO}, I_e, r$).

571 Simulation implementation

572 The modeled genome is always considered at 5 kb resolution. The input spatial profile $I_e + r$ is
573 normalised so that the sum is one. We draw from this normalised profile $N_{PO}(t = 0) = \frac{L}{d_{PO}}$ origins,
574 with replacement, meaning that several origins can be drawn in the same 5 kb window. During the
575 simulation we introduce $\rho_F L$ firing factors following an exponential characteristic law $\rho_F L(1 - e^{-t/\tau})$
576 with τ a characteristic time taken to 1 h to simulate progressive activation of firing factor upon
577 S-phase entry. We use a Gillespie algorithm (**Gillespie, 1976**) to simulate an initiation reaction with
578 reaction rate $k_{on} N_{PO}(t) N_F(t)$ with $N_F(t)$ the number of free firing factor at time t . The Gillespie algo-
579 rithm considers that the the next reaction between an origin and a firing factor will take place after
580 a time δt_A drawn from an exponential distribution of parameter $\frac{1}{k_{on} N_{PO}(t) N_F(t)}$. Then δt_A is compared
581 to δt_E , the smallest time of encounter for two forks on the genome. The system then evolve for
582 an increment of time $\delta t = \min(\delta t_E, \delta t_A)$, meaning that all the forks on the genome moves of $v\delta t$. If
583 $\delta t_A < \delta t_E$ then one origin is activated at random, a firing factor is trapped ($N_F(t + \delta t) = N_F(t) - 1$),
584 and two forks propagate on opposite directions at a velocity v from the origin. Otherwise the ter-
585 mination event releases a factor so that $N_F(t + \delta t) = N_F(t) + 1$ and a new time step begin. If a fork
586 replicates a position with unfired origins, then these origins are passivated and remove from the
587 list of potential origins.

588 Remark on rescaling the simulation for the simulation on chromosome 2 only.

589 In **Arbona et al. (2018)**, we showed that to reproduce the shape of the experimental temporal rate
590 of DNA replication origin firing, one has to be in a regime governed by the critical parameter $\rho_F^* =$
591 $\frac{v}{k_{on} L d_{PO}}$. In order to stay in this regime no matter the size of the genome, we in fact parameterized
592 the simulations with $k_{on}^e = k_{on} L$ chosen constant. $k_{on}^e = 8.625$ kb min⁻¹, so that when we simulated
593 the whole genome (in our case, the first 22 human chromosomes whose total size is 2 875 Mb), then
594 $k_{on} = 3e^{-6}$ min⁻¹. This means that if we decrease the size of the system, the constant of reaction k_{on}
595 increases, which is coherent as k_{on} encompass the efficiency of encounter between one potential
596 origin and one firing factor and being in a smaller system their encounter rate is increased.

597 Missing data.

598 For all simulations if a gap larger than 1.5 Mb without data in either MRT or RFD experimental data
599 was present, the region extended of 500 kb on both ends was removed. This mainly happen in
600 telomeric and centromeric regions, and it means that a chromosome can be segmented in two
601 or more pieces. Then when comparing e.g. simulated with experimental MRT, we also remove all
602 gaps in MRT data that are smaller than 1.5 Mb extended of 500 kb on both ends. These two steps
603 remove less than 10 % of the genome for either MRT or RFD in all considered cell lines. When
604 computing replication time, we exclude all gaps present in either MRT or RFD data as well their
605 surrounding 500 kb.

606 Computing experimental quantities, comparison with experimental data.

607 To compute RFD, we record for each 5 kb window in each simulation the fork direction as +1, -1
608 or 0 if an initiation or a termination occurred. The final RFD is the mean of 200 simulations. To
609 compute the MRT as done in repli-seq experiments, we recorded for each simulation and each
610 locus the actual replicated fraction of the genome at the time the locus is replicated. Then to
611 simulate the six fractions of the repli-seq experiment, the continuous [0..1] interval of replicated
612 fraction of the genome is mapped to six bin of length 1/6. Then $MRT(x) = \sum p_i(x) i/6 + 1/12$ where
613 $p_i(x)$ if the fraction of the simulations where the locus at position x has been replicated when the
614 replicated fraction of the genome was between $[i/6, (i + 1)/6]$ ($i \in \{0, \dots, 5\}$).

615 For both MRT and RFD when comparing with experimental data, we masked the region re-
616 moved as specified in the missing data paragraph. Pearson correlations were computed at 5 kb

617 resolution for RFD and 10 kb resolution for MRT. T_{100} was defined as the replication time of the
618 latest replicated window. T_{99} (resp. T_{95}) is defined as the time at which 99% (resp. 95%) of the
619 genome was replicated.

620 Experimental data

621 DNaseI HS data were downloaded from the ENCODE project (*Hansen et al., 2010; Thurman et al.,*
622 *2007*): (K562 DNaseI HS Uniform Peaks from ENCODE/Analysis (table wgEncodeAwgDnaseUwduke-
623 K562UniPk.narrowPeak).

624 ORC2 binding sites in K562 cells were obtained from *Miotto et al. (2016)* (supplementary mate-
625 rial table S1 in the *Miotto et al. (2016)* article).

626 SNS-seq data were obtained for K562 from *Picard et al. (2014)* (GSE46189_Ori-Peak.bed) and
627 for HeLa from *Besnard et al. (2012)* (Hela_SNS_seq_Besnard_Tot.1kb.csv).

628 Bubble-Seq data were obtained from *Mesner et al. (2013)* (GSE38809_GM_combined_RD_bubbles-
629 .bedgraph).

630 RFD profiles derived from OK-seq data were obtained for HeLa from *Petryk et al. (2016)* and for
631 GM06990 and K562 from *Wu et al. (2018)*.

632 For mean replication timing data, GM12878, K562, HeLaS3, alignment files of Repli-seq libraries
633 (BAM files) for six S-phase fractions were obtained from the ENCODE project (*Hansen et al., 2010;*
634 *Thurman et al., 2007*) at <http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliS>

635 For yeast, RFD profiles derived from OK-seq data were obtained from *Hennion et al. (2018)* and
636 MRT was obtained from *Müller et al. (2014)*. We shifted MRT by 0.05 for numerical stability.

637 Grid search optimisation

638 When performing grid search optimisation on the four parameters ρ_F , d_{PO} , r , v we noticed, as
639 explained in the main text, that both d_{PO} and v had little effect on the MRT and RFD profiles. These
640 two parameters were thus left out so that ρ_F and r optimisation was carried on a 2-dimensional
641 grid. The optimum selected is the one of highest sum of Pearson correlation between simulated
642 and experiment MRT and RFD. For the grid search optimisation carried on chromosome 2 (second
643 Results' section), the explored r values were [0,0.02,0.05,0.1], and ρ_F values were [0.6, 0.7, 0.8, 0.9,
644 1. , 1.1, 1.2, 1.3, 1.4] $\times\rho_{start}$ where ρ_{start} was defined as the firing factor density needed to replicate
645 the whole genome at a fork speed of 1.5 kb.min⁻¹ in an S-phase duration T_S , if all the firing factors
646 are active: $\rho_{start} = \frac{1}{2*v*T_S}$, with $T_S = 8$ h for HeLa cell and 12 h for GM06990 and K562.

647 When performed genome-wide (Results reported in Table 1), we chose the same grid for all
648 human cell lines. r varied from 0 to 20 % by increments of 5 %, and explored ρ_F values were [0.27
649 0.41 0.55 0.68 0.82 0.95 1.1] Mb⁻¹.

650 Iterative procedure used in learning the PODLS that best predicts both MRT 651 and RFD data.

652 To define a starting initiation profile, we computed the RFD increments between 5 kb bins, smoothed
653 the profile using a 50 kb average sliding window, kept values higher than the 80th percentile and
654 set the rest to zero. Using this profile we ran a grid search optimisation (See previous paragraph) to
655 obtain MRT_0 and RFD_0 that best fitted experimental data. Then a neural network N_1 was trained
656 to predict the initiation profile $I_0 + r$ (r being the amount of random activation obtained from the
657 grid search optimisation part) from the simulated MRT_0 and RFD_0 (See next paragraph for details
658 on the neural network). N_1 was then applied on experimental MRT and RFD predicting the I_1 pro-
659 file, which was then used in a simulation to obtain MRT_1 and RFD_1 and compute the correlation
660 with experimental MRT and RFD . We reiterated the process twice using successively MRT_1 and
661 RFD_1 as input and then the obtained MRT_2 and RFD_2 as input to produce MRT_3 and RFD_3 . We
662 reiterated the procedure once more and stopped as it did not improve the correlations. The code
663 to reproduce these steps is available at (<https://github.com/organic-chemistry/repli1D>).

664 Neural network training

665 The input of the network was a window of size 2005 kb (401 bins of 5 kb) with both MRT and RFD, and
666 the output was the initiation signal at the center of the window. The RFD was smoothed with a 50
667 kb rolling window. We used a three-layer convolutional neural network with kernel length 10 and
668 filter size 15 and a relu activation. Each layer was followed by a dropout layer at a value of 1%. The
669 last convolutional layer was followed by a maxpooling of kernel size 2. The resulting vector went
670 through a dense layer with sigmoid activation and output size 1, meaning that the 2005 kb window
671 allowed to compute the probability of activation at its center. We used a binary cross entropy loss
672 and the layer was trained with the adadelta algorithm. The procedure was implemented using
673 keras. We used chromosome 3 to 23 for the training, chromosome 2 for validation and chromo-
674 some 1 for testing. To make the network more robust to experimental noise, we randomly added
675 noise to the input RFD profile by assigning 1% of the bins a random value between -1 and 1.

676 Analytical extraction of n_e from ΔRFD and MRT data

677 Eq. (7) requires that ΔRFD be estimated. We smoothed human RFD data using a running average
678 window of 15 kb (3 bins) then computed ΔRFD between consecutive 5kb windows. The 15% top
679 values were selected and the other bins set to 0. The non-zero bins were divided by $1/MRT$ or by
680 e^{-6MRT} .

681 Selecting ΔRFD peaks

682 We ran a peak detection algorithm over ΔRFD at a 5 kb resolution, using scipy routine `find_peaks`
683 with parameters `width=4` and `height=0.02`, thus selecting peaks $\geq 4 \times 5$ kb in width and $\geq 0.02 \Delta RFD/5$ kb
684 in height. This yielded 9878 peaks for GM06990, 13466 peaks for HeLa and 10009 peaks for K562.
685 The selected peaks as well as ΔRFD and I_M for comparison are shown on Fig. S15.

References

- Arbona JM**, Goldar A, Hyrien O, Arneodo A, Audit B. The eukaryotic bell-shaped temporal rate of DNA replication origin firing emanates from a balance between origin activation and passivation. *eLife*. 2018 Jun; 7. doi: [10.7554/elife.35192](https://doi.org/10.7554/elife.35192).
- Audit B**, Baker A, Chen CL, Rappailles A, Guilbaud G, Julienne H, Goldar A, d'Aubenton Carafa Y, Hyrien O, Thermes C, Arneodo A. Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm. *Nature Protocols*. 2012 Dec; 8(1):98–110. doi: [10.1038/nprot.2012.145](https://doi.org/10.1038/nprot.2012.145).
- Baker A**, Bechhoefer J. Inferring the spatiotemporal DNA replication program from noisy data. *Physical Review E*. 2014; 89(3):032703.
- Baker A**, Audit B, Chen CL, Moindrot B, Leleu A, Guilbaud G, Rappailles A, Vaillant C, Goldar A, Mongelard F, d'Aubenton Carafa Y, Hyrien O, Thermes C, Arneodo A. Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. *PLoS Comput Biol*. 2012; 8(4):e1002443. <http://dx.doi.org/10.1371/journal.pcbi.1002443>, doi: [10.1371/journal.pcbi.1002443](https://doi.org/10.1371/journal.pcbi.1002443).
- Bazarova A**, Nieduszynski CA, Akerman I, Burroughs NJ. Bayesian inference of origin firing time distributions, origin interference and licencing probabilities from Next Generation Sequencing data. *Nucleic acids research*. 2019; 47(5):2229–2243.
- Bechhoefer J**, Rhind N. Replication timing and its emergence from stochastic processes. *Trends Genet*. 2012; 28(8):374–381. doi: [10.1016/j.tig.2012.03.011](https://doi.org/10.1016/j.tig.2012.03.011).
- Besnard E**, Babled A, Lapasset L, Milhavet O, Parrinello H, Dantec C, Marin JM, Lemaitre JM. Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs. *Nature Structural & Molecular Biology*. 2012 Jul; 19(8):837–844. doi: [10.1038/nsmb.2339](https://doi.org/10.1038/nsmb.2339).
- Blin M**, Lacroix L, Petryk N, Jaszczyszyn Y, Chen CL, Hyrien O, Le Tallec B. DNA molecular combing-based replication fork directionality profiling. *Nucleic Acids Research*. 2021; .
- Burkhart R**, Schulte D, Hu B, Musahl C, Göhring F, Knippers R. Interactions of human nuclear proteins P1Mcm3 and P1Cdc46. *European journal of biochemistry*. 1995; 228(2):431–438.

- Chen CL**, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, Audit B, d'Aubenton Carafa Y, Arneodo A, Hyrien O, et al. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome research*. 2010; 20(4):447–457.
- Collart C**, Allen GE, Bradshaw CR, Smith JC, Zegerman P. Titration of four replication factors is essential for the *Xenopus laevis* midblastula transition. *Science*. 2013; 341(6148):893–896.
- Conti C**, Sacca B, Herrick J, Lalou C, Pommier Y, Bensimon A. Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. *Mol Biol Cell*. 2007 Aug; 18(8):3059–3067. <http://dx.doi.org/10.1091/mbc.E06-08-0689>, doi: 10.1091/mbc.E06-08-0689.
- Das SP**, Borrmann T, Liu VWT, Yang SCH, Bechhoefer J, Rhind N. Replication timing is regulated by the number of MCMs loaded at origins. *Genome Res*. 2015 Dec; 25:1886–1892. doi: 10.1101/gr.195305.115.
- DePamphilis M**, Bell S. *Genome Duplication: concepts, mechanisms, evolution and disease*. Garland Science, New York; 2010.
- Douglas ME**, Ali FA, Costa A, Diffley JF. The mechanism of eukaryotic CMG helicase activation. *Nature*. 2018; 555(7695):265–268.
- Douglas ME**, Diffley JF. Replication timing: the early bird catches the worm. *Current Biology*. 2012; 22(3):R81–R82.
- Edwards MC**, Tutter AV, Cvetic C, Gilbert CH, Prokhorova TA, Walter JC. MCM2–7 Complexes Bind Chromatin in a Distributed Pattern Surrounding the Origin Recognition Complex in *Xenopus* Egg Extracts. *Journal of Biological Chemistry*. 2002; 277(36):33049–33057.
- Evrin C**, Clarke P, Zech J, Lurz R, Sun J, Uhle S, Li H, Stillman B, Speck C. A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. *Proceedings of the National Academy of Sciences*. 2009; 106(48):20240–20245.
- Foss EJ**, Sripathy S, Gatbonton-Schwager T, Kwak H, Thiesen AH, Lao U, Bedalov A. Chromosomal Mcm2-7 distribution and the genome replication program in species from yeast to humans. *PLoS Genetics*. 2021; 17(9):e1009714.
- Gillespie DT**. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comput Phys*. 1976; 22(4):403–434. <http://www.sciencedirect.com/science/article/pii/0021999176900413>, doi: 10.1016/0021-9991(76)90041-3.
- Gindin Y**, Valenzuela MS, Aladjem MI, Meltzer PS, Bilke S. A chromatin structure-based model accurately predicts DNA replication timing in human cells. *Molecular Systems Biology*. 2014 3; 10(3):722–722. doi: 10.1002/msb.134859.
- Goldar A**, Labit H, Marheineke K, Hyrien O. A dynamic stochastic model for DNA replication initiation in early embryos. *PLoS One*. 2008; 3(8):e2919. <http://dx.doi.org/10.1371/journal.pone.0002919>, doi: 10.1371/journal.pone.0002919.
- Goldar A**, Marsolier-Kergoat MC, Hyrien O. Universal temporal profile of replication origin activation in eukaryotes. *PLoS One*. 2009; 4(6):e5899. <http://dx.doi.org/10.1371/journal.pone.0005899>, doi: 10.1371/journal.pone.0005899.
- Guilbaud G**, Rappailles A, Baker A, Chen CL, Arneodo A, Goldar A, d'Aubenton Carafa Y, Thermes C, Audit B, Hyrien O. Evidence for Sequential and Increasing Activation of Replication Origins along Replication Timing Gradients in the Human Genome. *PLoS Computational Biology*. 2011 Dec; 7(12):e1002322. doi: 10.1371/journal.pcbi.1002322.
- Hahn AT**, Jones JT, Meyer T. Quantitative analysis of cell cycle phase durations and PC12 differentiation using fluorescent biosensors. *Cell cycle*. 2009; 8(7):1044–1052.
- Hamlin JL**, Mesner LD, Dijkwel PA. A winding road to origin discovery. *Chromosome research*. 2010; 18(1):45–61.
- Hansen RS**, Thomas S, Sandstrom R, Canfield TK, Thurman RE, Weaver M, Dorschner MO, Gartler SM, Stamatoyanopoulos JA. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*. 2010; 107(1):139–144.

- Harvey KJ**, Newport J. CpG methylation of DNA restricts prereplication complex assembly in *Xenopus* egg extracts. *Molecular and cellular biology*. 2003; 23(19):6769–6779.
- Hennion M**, Arbona JM, Cruaud C, Proux F, Le Tallec B, Novikova E, Engelen S, Lemainque A, Audit B, Hyrien O. Mapping DNA replication with nanopore sequencing. *bioRxiv*. 2018; <https://www.biorxiv.org/content/early/2018/09/26/426858>, doi: 10.1101/426858.
- Hennion M**, Arbona JM, Lacroix L, Cruaud C, Theulot B, Le Tallec B, Proux F, Wu X, Novikova E, Engelen S, et al. FORK-seq: replication landscape of the *Saccharomyces cerevisiae* genome by nanopore sequencing. *Genome biology*. 2020; 21(1):1–25.
- Hulke ML**, Massey DJ, Koren A. Genomic methods for measuring DNA replication dynamics. *Chromosome Research*. 2020; 28(1):49–67.
- Hyrien O**. Peaks cloaked in the mist: the landscape of mammalian replication origins. *Journal of Cell Biology*. 2015; 208(2):147–160.
- Ibarra A**, Schwob E, Méndez J. Excess MCM proteins protect human cells from replicative stress by licensing backup origins of replication. *Proceedings of the National Academy of Sciences*. 2008; 105(26):8956–8961.
- Kirstein N**, Buschle A, Wu X, Krebs S, Blum H, Kremmer E, Vorberg IM, Hammerschmidt W, Lacroix L, Hyrien O, et al. Human ORC/MCM density is low in active genes and correlates with replication time but does not delimit initiation zones. *Elife*. 2021; 10:e62161.
- Koren A**, Handsaker RE, Kamitaki N, Karlić R, Ghosh S, Polak P, Eggen K, McCarroll SA. Genetic variation in human DNA replication timing. *Cell*. 2014; 159(5):1015–1026.
- Kumagai A**, Dunphy WG. Binding of the Treslin-MTBP complex to specific regions of the human genome promotes the initiation of DNA replication. *Cell reports*. 2020; 32(12):108178.
- Letessier A**, Millot GA, Koundrioukoff S, Lachagès AM, Vogt N, Hansen RS, Malfoy B, Brison O, Debatisse M. Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. *Nature*. 2011; 470(7332):120–123.
- Löb D**, Lengert N, Chagin VO, Reinhart M, Casas-Delucchi CS, Cardoso MC, Drossel B. 3D replicon distributions arise from stochastic initiation and domino-like DNA replication progression. *Nat Commun*. 2016 Apr; 7:11207. doi: 10.1038/ncomms11207.
- Long H**, Zhang L, Lv M, Wen Z, Zhang W, Chen X, Zhang P, Li T, Chang L, Jin C, et al. H2A. Z facilitates licensing and activation of early replication origins. *Nature*. 2020; 577(7791):576–581.
- Mantiero D**, Mackenzie A, Donaldson A, Zegerman P. Limiting replication initiation factors execute the temporal programme of origin firing in budding yeast. *The EMBO journal*. 2011; 30(23):4805–4814.
- Mei L**, Kedziora KM, Song EA, Purvis JE, Cook JG. The consequences of differential origin licensing dynamics in distinct chromatin environments. *bioRxiv*. 2021; .
- Mesner LD**, Valsakumar V, Ciešlik M, Pickin R, Hamlin JL, Bekiranov S. Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early-and late-firing origins. *Genome research*. 2013; 23(11):1774–1788.
- Miller TC**, Locke J, Greiwe JF, Diffley JF, Costa A. Mechanism of head-to-head MCM double-hexamer formation revealed by cryo-EM. *Nature*. 2019; 575(7784):704–710.
- Miotto B**, Ji Z, Struhl K. Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. *Proceedings of the National Academy of Sciences*. 2016 Aug; 113(33):E4810–E4819. <http://www.pnas.org/content/113/33/E4810>, doi: 10.1073/pnas.1609060113.
- de Moura AP**, Retkute R, Hawkins M, Nieduszynski CA. Mathematical modelling of whole chromosome replication. *Nucleic Acids Res*. 2010; 38(17):5623–5633.
- Müller CA**, Boemo MA, Spingardi P, Kessler BM, Kriaucionis S, Simpson JT, Nieduszynski CA. Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads. *Nature methods*. 2019; 16(5):429–436.
- Müller CA**, Hawkins M, Retkute R, Malla S, Wilson R, Blythe MJ, Nakato R, Komata M, Shirahige K, de Moura AP, et al. The dynamics of genome replication using deep sequencing. *Nucleic Acids Research*. 2014; 42(1):e3–e3.

- Pereira PD**, Serra-Caetano A, Cabrita M, Bekman E, Braga J, Rino J, Santus R, Filipe PL, Sousa AE, Ferreira JA. Quantification of cell cycle kinetics by EdU (5-ethynyl-2-deoxyuridine)-coupled-fluorescence-intensity analysis. *Oncotarget*. 2017; 8(25):40514.
- Petryk N**, Dalby M, Wenger A, Stromme CB, Strandsby A, Andersson R, Groth A. MCM2 promotes symmetric inheritance of modified histones during DNA replication. *Science*. 2018; 361(6409):1389–1392.
- Petryk N**, Kahli M, d'Aubenton Carafa Y, Jaszczyszyn Y, Shen Y, Silvain M, Thermes C, Chen CL, Hyrien O. Replication landscape of the human genome. *Nat Commun*. 2016 Jan; 7:10208. <http://www.nature.com/ncomms/2016/160111/ncomms10208/full/ncomms10208.html>, doi: 10.1038/ncomms10208.
- Picard F**, Cadoret JC, Audit B, Arneodo A, Alberti A, Battail C, Duret L, Prioleau MN. The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells. *PLoS Genetics*. 2014 5; 10(5):e1004282. doi: 10.1371/journal.pgen.1004282.
- Remus D**, Beuron F, Tolun G, Griffith JD, Morris EP, Diffley JF. Concerted loading of Mcm2–7 double hexamers around DNA during DNA replication origin licensing. *Cell*. 2009; 139(4):719–730.
- Schmidt JM**, Bleichert F. Structural mechanism for replication origin binding and remodeling by a metazoan origin recognition complex and its co-loader Cdc6. *Nature communications*. 2020; 11(1):1–17.
- Siddiqui K**, On KF, Diffley JF. Regulating DNA replication in eukarya. *Cold Spring Harbor perspectives in biology*. 2013; 5(9):a012930.
- Sugimoto N**, Maehara K, Yoshida K, Ohkawa Y, Fujita M. Genome-wide analysis of the spatiotemporal regulation of firing and dormant replication origins in human cells. *Nucleic acids research*. 2018; 46(13):6683–6696.
- Tanaka S**, Nakato R, Katou Y, Shirahige K, Araki H. Origin association of Sld3, Sld7, and Cdc45 proteins is a key step for determination of origin-firing timing. *Current Biology*. 2011; 21(24):2055–2063.
- Técher H**, Koundrioukoff S, Azar D, Wilhelm T, Carignon S, Brison O, Debatisse M, Le Tallec B. Replication dynamics: biases and robustness of DNA fiber analysis. *Journal of molecular biology*. 2013; 425(23):4845–4855.
- Thurman RE**, Day N, Noble WS, Stamatoyannopoulos JA. Identification of higher-order functional domains in the human ENCODE regions. *Genome research*. 2007; 17(6):917–927.
- Tubbs A**, Sridharan S, van Wietmarschen N, Maman Y, Callen E, Stanlie A, Wu W, Wu X, Day A, Wong N, et al. Dual roles of poly (dA: dT) tracts in replication initiation and fork collapse. *Cell*. 2018; 174(5):1127–1142.
- Wang W**, Klein KN, Proesmans K, Yang H, Marchal C, Zhu X, Borrmann T, Hastie A, Weng Z, Bechhoefer J, et al. Genome-wide mapping of human DNA replication by optical replication mapping supports a stochastic model of eukaryotic replication. *Molecular Cell*. 2021; .
- Weber TS**, Jaehnert I, Schichor C, Or-Guil M, Carneiro J. Quantifying the length and variance of the eukaryotic cell cycle phases by a stochastic model and dual nucleoside pulse labelling. *PLoS computational biology*. 2014; 10(7):e1003616.
- Weis MC**. Computational models of the mammalian cell cycle. Case Western Reserve University; 2012.
- Widrow R**, Hansen RS, Kawame H, Gartler SM, Laird CD. Very late DNA replication in the human cell cycle. *Proceedings of the National Academy of Sciences*. 1998; 95(19):11246–11250.
- Wittig KA**, Sansam CG, Noble TD, Goins D, Sansam CL. The CRL4DTL E3 ligase induces degradation of the DNA replication initiation factor TICRR/TRESLIN specifically during S phase. *Nucleic acids research*. 2021; 49(18):10507–10523.
- Wong PG**, Winter SL, Zaika E, Cao TV, Oguz U, Koomen JM, Hamlin JL, Alexandrow MG. Cdc45 limits replicon usage from a low density of preRCs in mammalian cells. *PLoS one*. 2011; 6(3):e17533.
- Wu X**, Kabalane H, Kahli M, Petryk N, Laperrousaz B, Jaszczyszyn Y, Drillon G, Nicolini FE, Perot G, Robert A, Fund C, Chibon F, Xia R, Wiels J, Argoul F, Maguer-Satta V, Arneodo A, Audit B, Hyrien O. Developmental and cancer-associated plasticity of DNA replication preferentially targets GC-poor, lowly expressed and late-replicating regions. *Nucleic Acids Research*. 2018 09; 46(19):10157–10172. <https://doi.org/10.1093/nar/gky797>, doi: 10.1093/nar/gky797.

Yang SCH, Rhind N, Bechhoefer J. Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing. *Mol Syst Biol.* 2010; 6(1):404.

Zhao PA, Sasaki T, Gilbert DM. High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome biology.* 2020; 21(1):1–20.

Supporting information

PODLS	K562 MRT	K562 RFD	GM MRT	GM RFD	Hela MRT	Hela RFD	<i>S. cer.</i> MRT	<i>S. cer.</i> RFD
I_0	0.81	0.79	0.74	0.77	0.78	0.74	0.60	0.71
I_1	0.93	0.89	0.95	0.88	0.95	0.82	0.80	0.83
I_2	0.98	0.91	0.98	0.91	0.97	0.85	0.87	0.84
I_3	0.98	0.92	0.99	0.91	0.98	0.85	0.94	0.90
I_4	0.98	0.92	0.99	0.91	0.99	0.84	0.96	0.91

Table S1. Pearson correlation coefficients between experimental MRT and RFD profiles and their simulated MRT_n and RFD_n estimates obtained for the series of iteratively optimized PODLS I_n using RFD derivative for initialisation of I_0 . Results are shown for K562, GM and Hela cell lines as well as *S. cerevisiae*. At the 5th iteration none of the PCC increased (not shown).

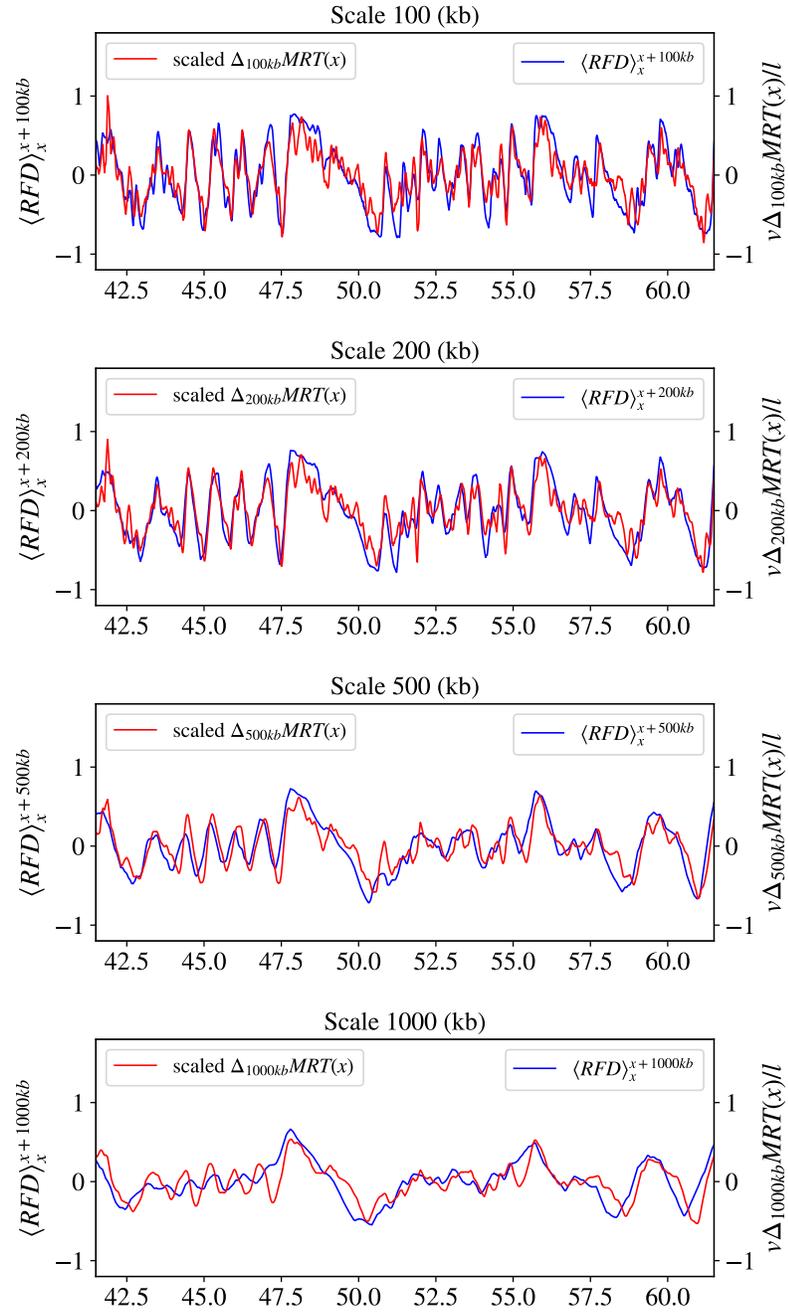


Figure S1. Comparison between $\langle \text{RFD} \rangle_x^{x+l}$ (blue) and $\frac{v}{l} \Delta_l \text{MRT}(x)$ (red) (Eq. (2)) assuming $T_S = 12\text{h}$ at different scale l . From (top) to (bottom), $l = 100\text{ kb}, 200\text{ kb}, 500\text{ kb}$ and 1000 kb and fork speed values v are taken from Fig. 2B. The same 20 Mb region of chromosome 1 as in Fig. 2A is shown.

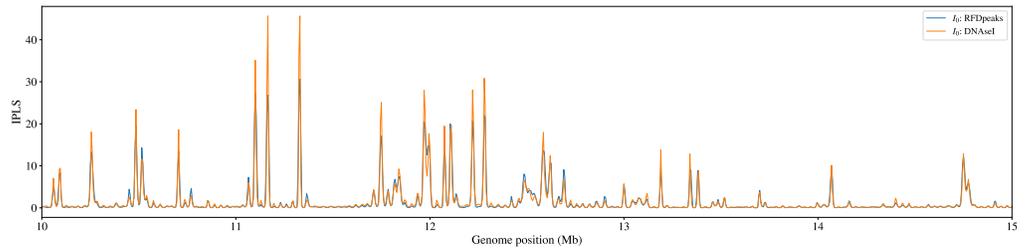


Figure S2. Comparison of K562 potential origin density landscapes optimized either from the PODLS derived from DNase I HS peaks (orange) or from the experimental ΔRFD peaks (blue). The obtained I_M are almost identical (PCC = 0.94).

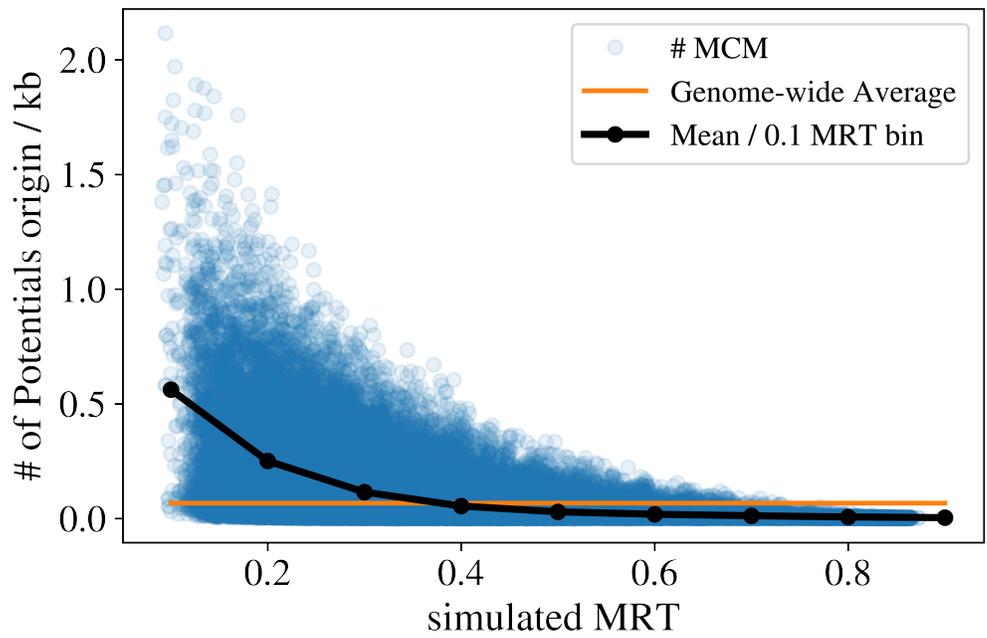


Figure S3. Predicted number of potential origins per kb, computed for each 5 kb bin, vs. MRT in K562 replication simulations.

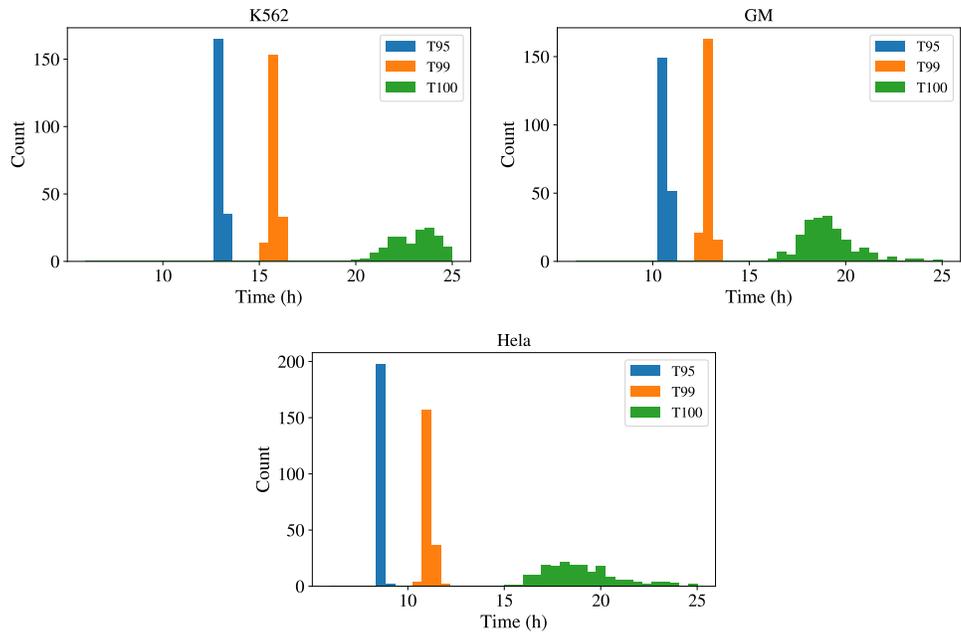


Figure S4. Distribution of T95 (blue), T99 (orange) and T100 (green) replication times as defined in the text for the three different cell lines.

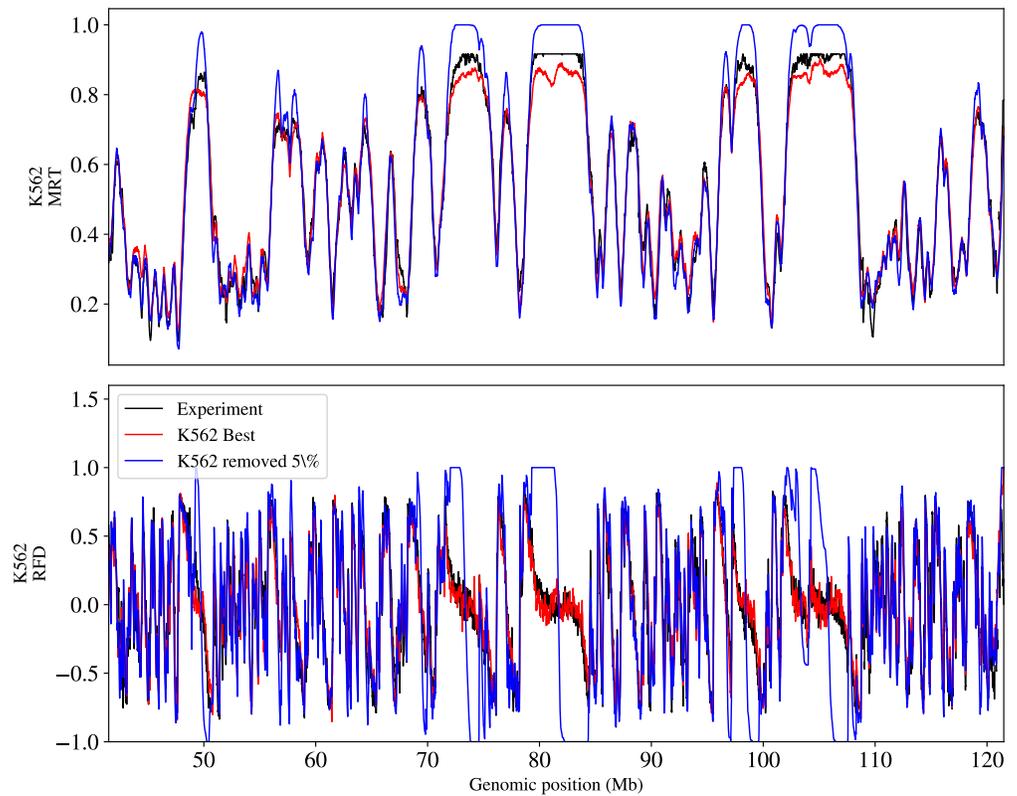


Figure S5. (Top) Comparison of experimental MRT with simulated MRT using the optimized PODLS I_M (red) and after setting to zero the bins with the lowest I_M values corresponding to 5% of the total origin firing events ($\approx 53\%$ of the bins) (blue). (Bottom) Same comparison for the RFD profiles.

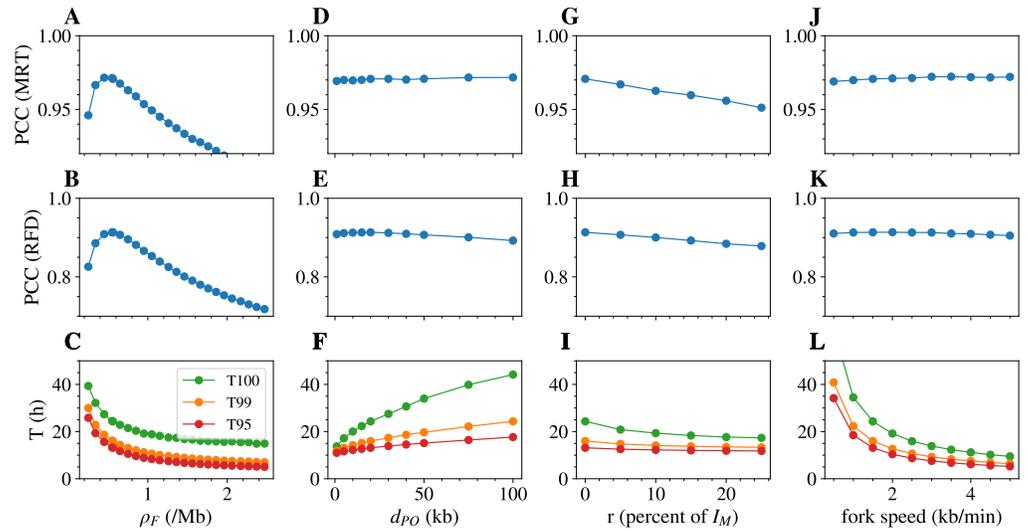


Figure S6. Effect of single parameter variation on measurable targets in GM06990. Effect of the density of firing factors ρ_F (A,B,C); the average distance between potential origins d_{PO} (D,E,F); the percent of random initiation r (G,H,I); the fork speed v (J,K,L), on the Pearson Correlation Coefficient (PCC) between simulated and experimental MRT (A,D,G,J) and RFD (B,E,H,K) profiles, and on T95 (red), T99 (orange) and T100 (green), the median times required to replicate 99% and 100% of the genome (C,F,I,L).

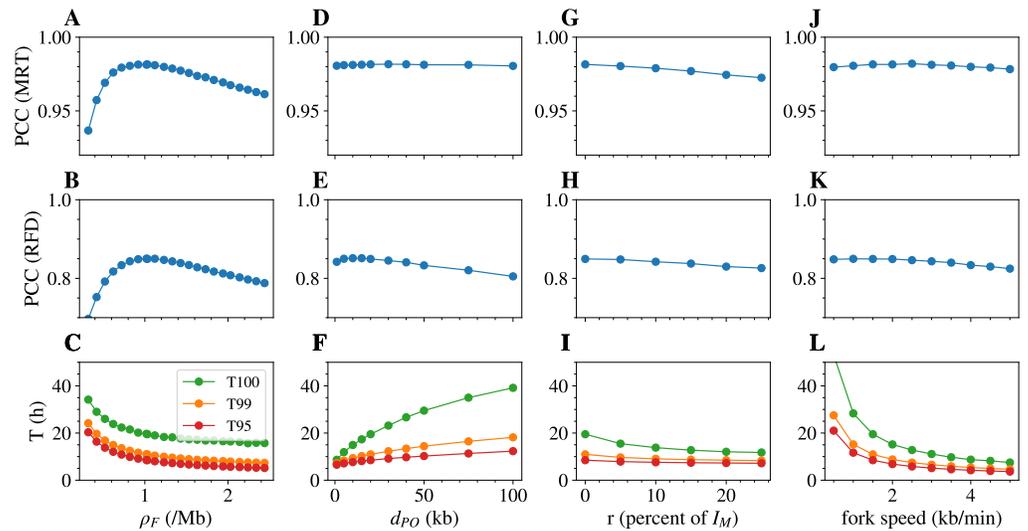


Figure S7. Effect of single parameter variation on measurable targets in HeLa. Effect of the density of firing factors ρ_F (A,B,C); the average distance between potential origins d_{PO} (D,E,F); the percent of random initiation r (G,H,I); the fork speed v (J,K,L), on the Pearson Correlation Coefficient (PCC) between simulated and experimental MRT (A,D,G,J) and RFD (B,E,H,K) profiles, and on n T95 (red), T99 (orange) and T100 (green), the median times required to replicate 99% and 100% of the genome (C,F,I,L).

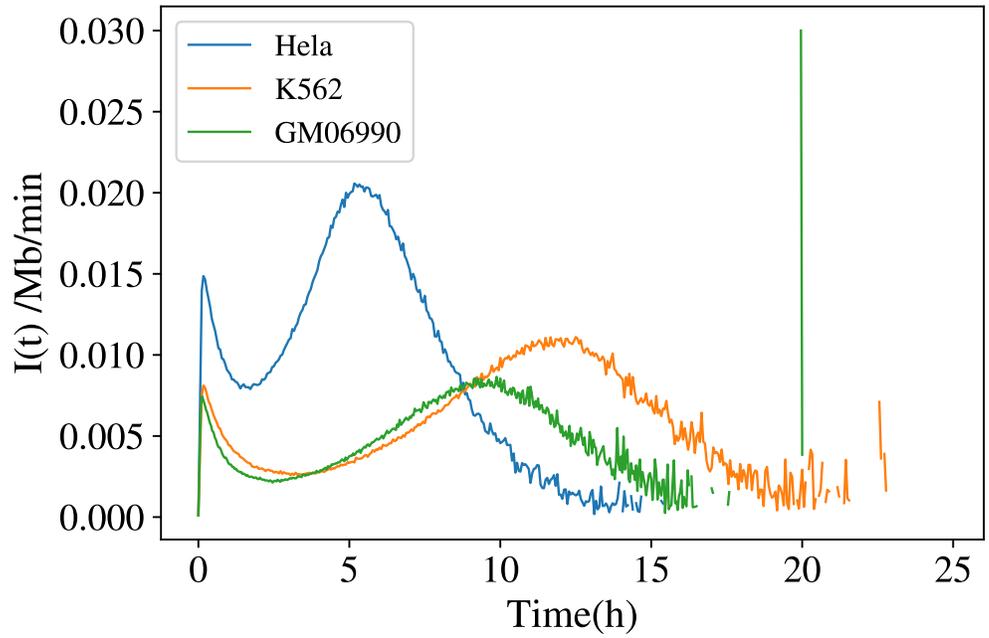


Figure S8. Probability of initiation per length of unreplicated DNA per minute for the three indicated cell lines

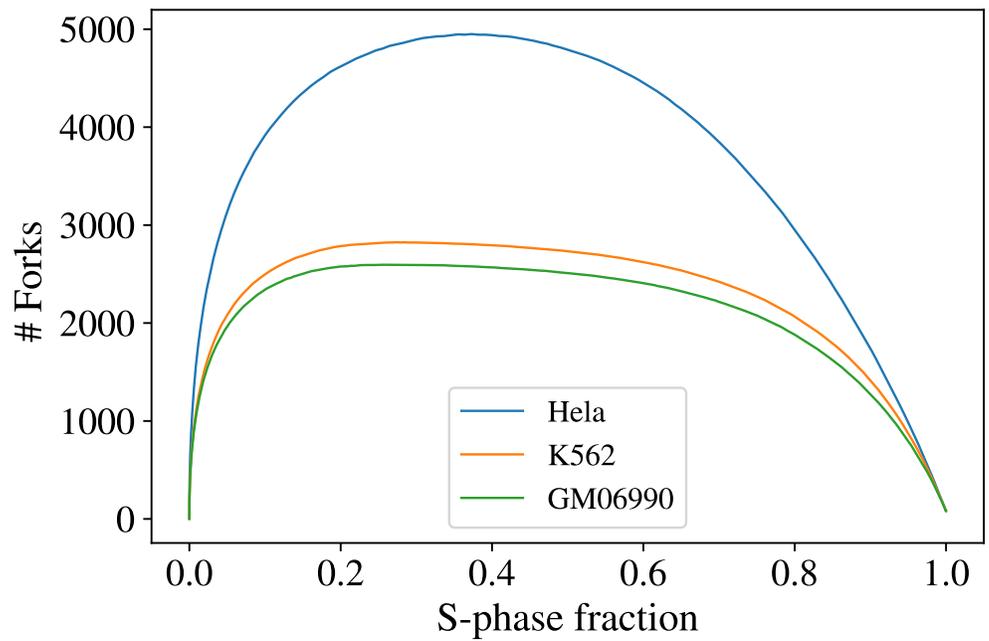


Figure S9. Active fork number as a function of S-phase fraction for the three indicated cell lines

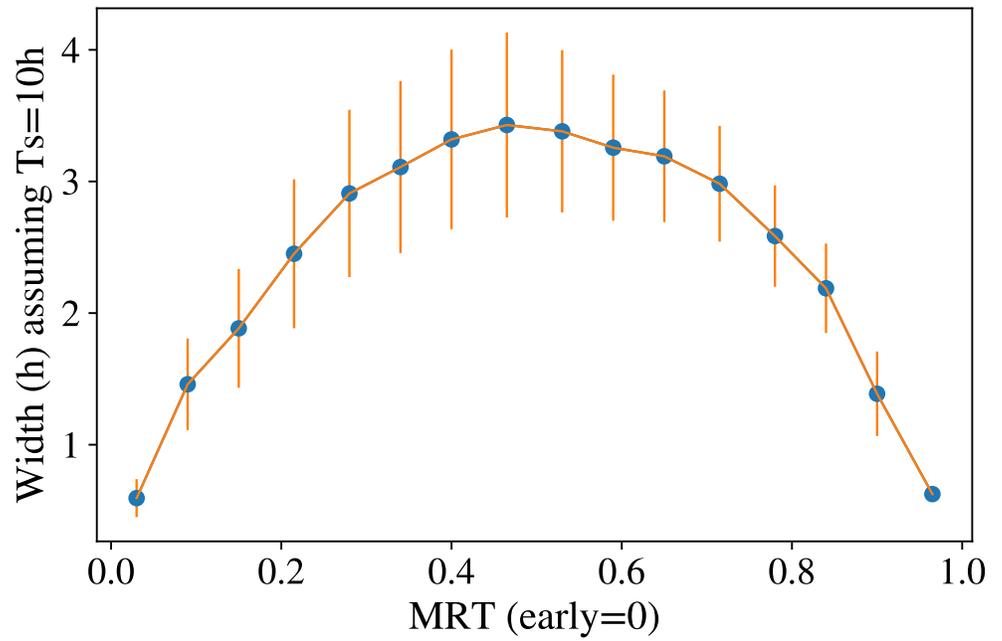


Figure S10. Replication time variability as a function of MRT in simulated K562 replication. Blue dots and orange bars indicate the genome-wide average and range of values, respectively of replication time variability.

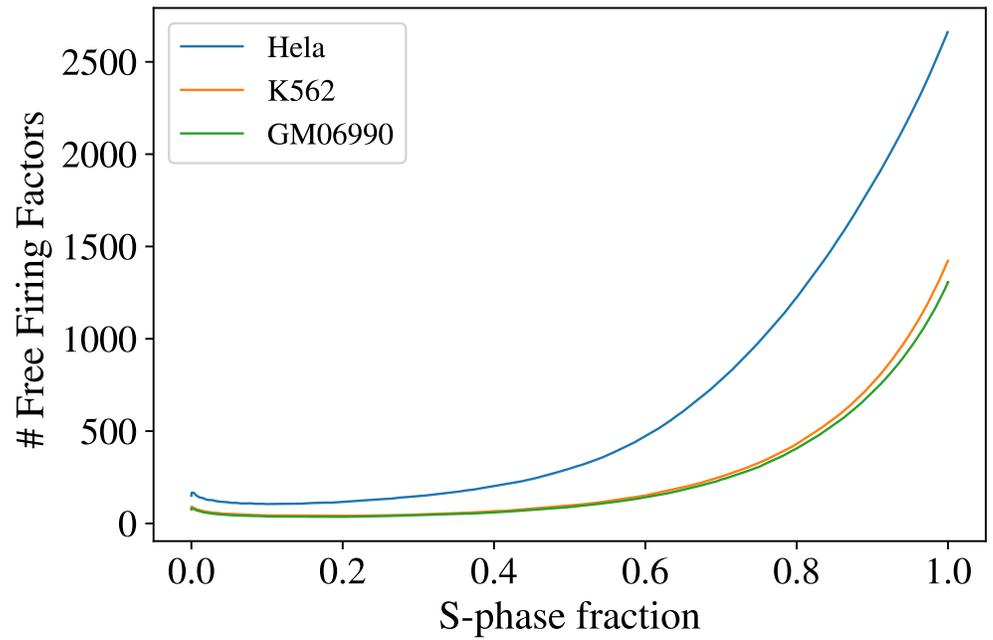


Figure S11. Number of Free firing factors as a function of the S-phase fraction for the three indicated cell lines

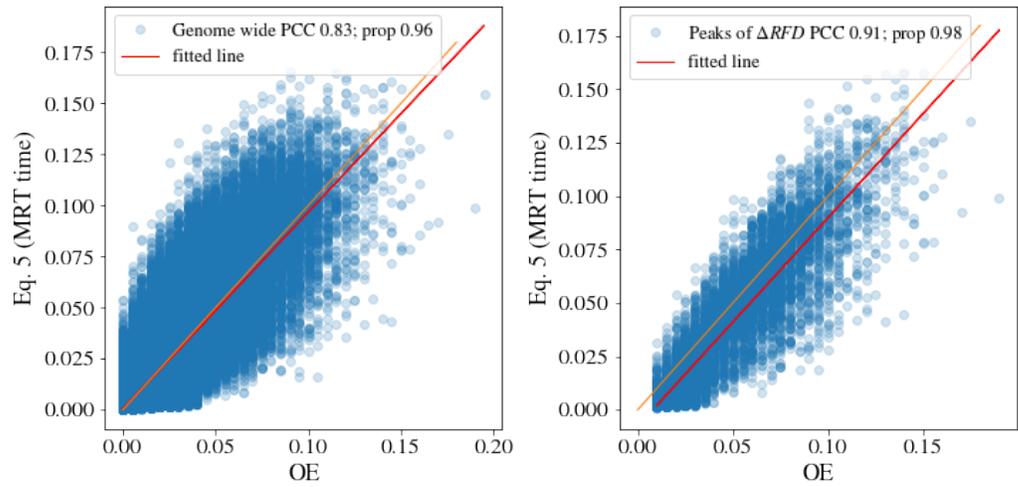


Figure S12. Comparison of observed origin efficiency in K562 replication simulation, directly counted as the fraction of simulations in which replication started in a bin (OE), or computed as the right-side term of Eq (5), genome wide (left) or restricted to the peaks of ΔRFD (right). Red line represents the linear fit and the orange line the first diagonal.

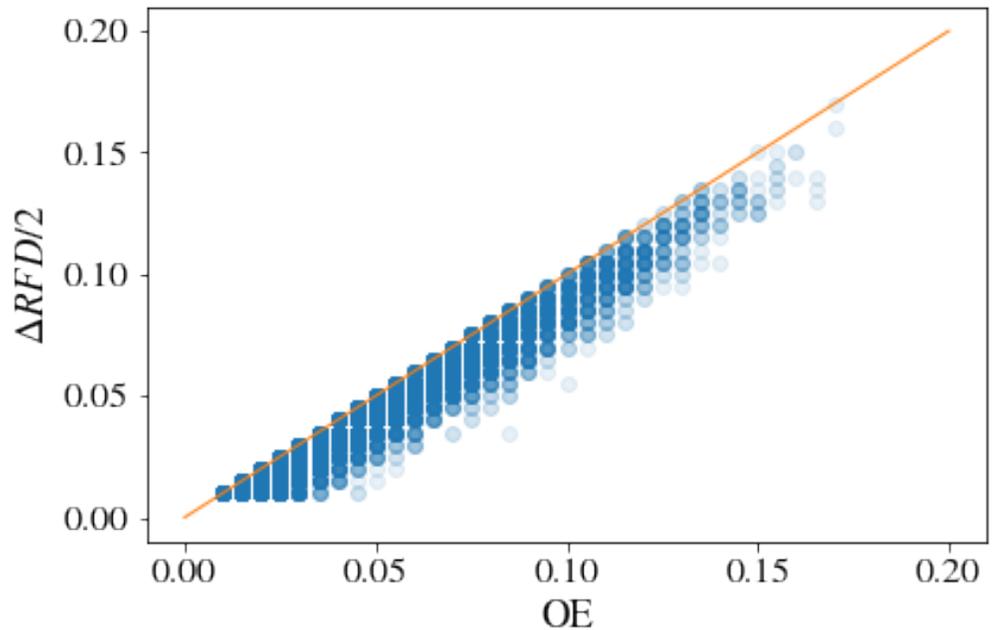


Figure S13. Comparison of OE, directly counted as the fraction of simulations in which replication started in a bin (OE), with its estimation by $\Delta RFD/2$ restricted to peaks of ΔRFD . The orange line is the first diagonal.

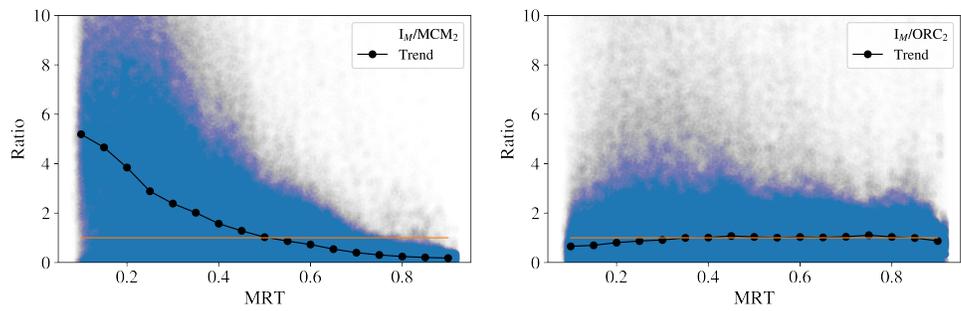


Figure S14. I_M/MCM_2 ratio in HeLa (left). I_M/ORC_2 ratio in K562 (right). Both signals were smoothed with a 50 kb sliding window due to noise in MCM data and normalized so that the median value over all the genome was one (orange line). The black dotted lines indicate the median value of the ratios by 0.05 MRT steps.

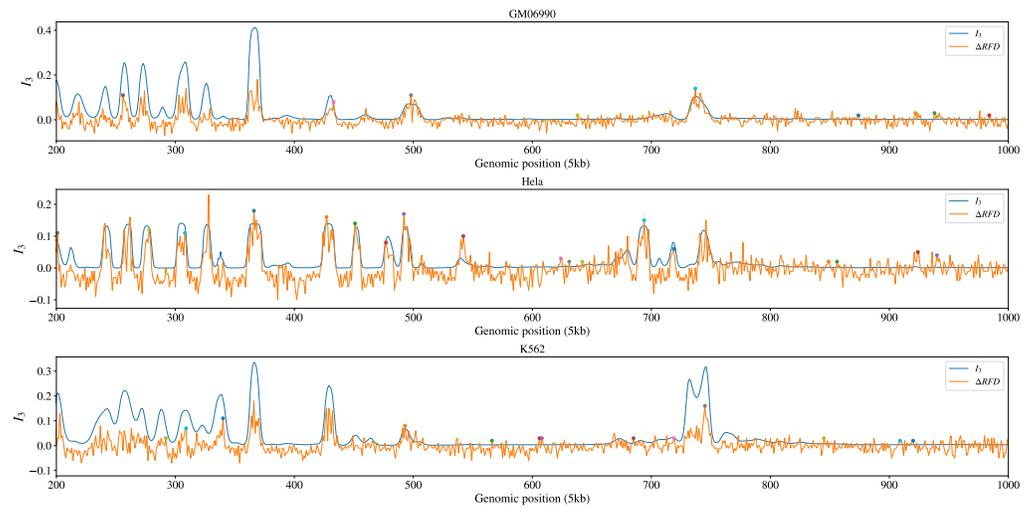


Figure S15. Peak detection of ΔRFD (dots on the top of the orange signal) overlaid with I_M (blue curves) for the three indicated cell lines