Inferring the DNA replication origin density landscape in human cells from replication timing and fork directionality data

Jean-Michel Arbona[1,*], Hadi Kabalane [2,¤], Arach Goldar[3], Olivier Hyrien [4,*], Benjamin Audit[2,*]

1 Laboratoire de Biologie et Modélisation de la Cellule, ENS de Lyon, Lyon, France
2 ENS de Lyon, CNRS, Laboratoire de Physique, F-69342 Lyon, France
3 Ibitec-S, CEA, France
4 Ecole Normale Supérieure, CNRS, INSERM, PSL Research University, France

¤Current Address: Centre de Recherches en Cancérologie de Toulouse, Inserm, Université Paul Sabatier, CNRS, Toulouse, France.
* jeanmichel.arbona@ens-lyon.fr; hyrien@bio.ens.psl.eu; benjamin.audit@ens-lyon.fr

Abstract

The determinants of the locations and firing times of the multiple replication origins are still elusive in human and other metazoan organisms. Experiments can independently profile mean replication timing (MRT) and replication fork directionality (RFD) genome-wide. In the hypothesis of a constant replication fork speed, MRT and RFD are related to each other by an analytical formula so are a priori equivalent. However, we show here that experimental noises result in MRT and RFD profiles containing information at different spatial frequencies. We further demonstrate that one can compute an origin density landscape that, when inserted in an appropriate simulation framework, jointly predicts experimental MRT and RFD profiles with an unprecedented precision. We also extract an analytical formula linking intrinsic origin efficiency with observed origin efficiency and MRT. We then compare the computed origin density landscape with experimental distributions of potential origins (ORC, MCM) or actual initiation events (Bubble-seq, SNS-seq, OK-seq). The results indicate that MRT and RFD data are highly consistent with each other, that our simple model suffices to capture the replication dynamics during S phase given an appropriate initiation probability landscape, but that the density of potential origins is not the sole determinant of this landscape.

Introduction

In eukaryotes, DNA replication starts at multiple chromosomal sites referred to as replication origins [12]. During the G1 phase of the cell cycle, origins are licensed by loading two MCM2-7 replicative helicase in the form of an inactive, head-to-head double hexamer (DH) ring around origin DNA [16, 47, 38, 48]. This symmetric configuration prepares the helicases to initiate bidirectional replication upon activation. Origin activation (or firing) can take place at different times through S phase, by binding of multiple firing factors that trigger origin DNA unwinding, conversion of the two inactive MCM2-7 rings into two active

Cdc45/MCM/GINS (CMG) helicases, each encircling and translocating 3'-to-5' along a single DNA strand, followed by recruitment of other replisome components [14]. Only a fraction of MCM DHs lead to productive initiation events, while the rest is inactivated by passing replication from other origins. This inactivation mechanism, called origin passivation [1], in addition to the mechanisms that restrict MCM2-7 loading to G1 phase, ensures that no part of the genome is replicated more than once during each cell cycle (Siddiqui et al 2013).

The mechanisms that determine mammalian origin positions and firing times are incompletely understood [30, 29]. The density of potential origins along the genome can be estimated by experimental detection of MCM [17, 31] and ORC [31, 39] proteins, ORC being the protein complex that loads MCMs on DNA. ORC and MCM profiles are indeed highly, though not perfectly, concordant [31]. In contrast to potential origins, the density of actual initiation events can be estimated by sequencing purified replication initiation intermediates, such as short nascent DNA strands (SNS-Seq [46]) or bubble-containing restriction fragments (Bubble-Seq [37]). However, these two methods are only weakly concordant [30, 29]. Other methods, based on sequencing elongation rather than initiation intermediates, reveal the average direction or time of replication through the entire genome, allowing to infer the location of initiation and termination events. Thus, mean replication timing (MRT) profiles have been computed by sequencing newly replicated DNA from sorted cells at different stages of S phase (Repli-seq; [9, 25, 61]) or by determining DNA copy number from proliferating cells [32]. Peaks of early MRT must contain origins, but the low resolution of MRT in human cells (50-100 kb [9, 25, 61]) has long precluded precise origin mapping. Replication fork directionality (RFD) profiles of human cells, obtained by strand-oriented sequencing of purified Okazaki fragments (OK-seq) were more resolutive (< 5 kb) [45, 59]. They revealed in each cell line 5,000 - 10,000 broad (10-100 kb) initiation zones (IZs), often flanking active genes, across which fork direction progressively shifted from left to right. They also revealed extended regions of unidirectional replication flanking IZs, fork merging zones between IZs, and randomly replicating regions in silent heterochromatin. Importantly, OK-seq IZs were confirmed genome-wide by EdUseq-HU [53], high-resolution Repli-Seq [61] and Optical Replication Mapping [54]. Initiation events may additionally occur outside IZs, but in a too dispersed manner to be directly detected in cell population profiles. Recent single-molecule and OK-seq analyses of the yeast genome [41, 27] and a model chicken locus [8] provided direct evidence for dispersed initiation in these systems. IZs can be shared between cell types or specific to a cell type. They are enriched in DNAse I hypersensitive sites (HSSs) and histone modifications such as H3K4me1, H3K27ac and H2A.Z, that usually mark active transcriptional regulatory elements [45, 59, 44]. H2A.Z was proposed to facilitate origin licensing and firing by recruiting SUV420H1, which promotes H4K20me2 deposition, in turn facilitating ORC binding [35].

There is now ample evidence that reproducible spatial structure of genomic replication profiles can emerge from stochastic firing of individual origins [6]. Modeling studies showed that a probabilistic interaction of potential origins with rate-limiting firing factors, engaged with forks and recycled at termination events, can predict the time-dependent profiles of origin firing rate and fork density universally observed in eukaryotes [1, 21, 20]. Experimental studies indeed suggested that rate-limiting activators regulate replication kinetics in yeast [36, 50] and metazoans [58, 10]. Thus, a simple model for replication regulation is that potential origins fire at different mean times because of their different affinities for limiting factors [13]. Alternatively, potential origins may all have the same affinity for firing factors but their variable density along the genome may determine MRT [60, 11]. We refer to these distinct

models as the origin affinity model and the origin density model, respectively.

Gindin et al. [19] built a kinetic model in which the time-dependent probability of initiation at a yet unreplicated site was the product of the time-dependent availability of a limiting factor by the time-independent, local value of a genomic "initiation probability landscape" (IPLS). Of the various genomic and epigenomic profiles used as estimate of IPLS, DNase I HSS profiles produced the best match with experimental MRT profiles (Pearson correlation between simulated and experimental MRT of 0.865). Importantly, the same IPLSs did not produce realistic MRT profiles in models that did not include competition for limiting fork-associating factors [19]. Since the Gindin model did not explicitly separate origin licensing and firing, however, it remained unclear whether the IPLS reflected potential origin density, or affinity, or both. ORC and MCM abundances, presumably the best direct indicators of potential origin density, correlated well with DNase I HSS [39] and early MRT [17, 31]. ORC- or MCM-based IPLSs produced realistic MRT profiles in Gindin-like simulations [39, 17]. However, our comparison of ORC, MCM and RFD profiles of the Raji cell line showed that when confounding parameters such as MRT and transcription status are controlled, ORC and MCM densities are not predictive of IZs. This suggested that potential origins may be more widespread than initiation sites but have different firing efficiencies, perhaps due to specific MCM or histone modifications affecting their affinities for firing factors [31].

Here, we ask whether the correlation of DNase I HSS with origin activation seen at MRT resolution (50-100 kb) still holds true at RFD resolution (< 5 kb). Although we find a remarkable conformity of MRT and RFD data to a simple mathematical equation that links both profiles, we demonstrate that MRT and RFD provide information at different scales. We explore their joint predictability using an elaboration of our previous kinetic model of DNA replication [1] in which discrete, localized potential origins (i.e. MCM DHs), chosen from an arbitrary potential origin density landscape (PODLS), are activated in a probabilistic manner by interaction with limiting firing factors that engage with forks and are released at termination (Fig. 1). As each potential origin is given the same probability to fire, the non-uniformity of the obtained replication profiles only comes from the non-uniformity of the PODLS (origin density model). We test PODLSs based on DNase I HSS, ORC, MCM, Bubble-seq and SNS-seq data. Furthermore, we train a neural network on simulated MRT and RFD profiles to infer a PODLS that predicts experimental MRT and RFD almost exactly, surpassing any PODLS based on experimental data. Then, we compare this optimized PODLS to SNS-seq, bubble-seq and OK-seq data (actual initiation events), using a novel mathematical expression that relates observed origin efficiencies and PODLS through MRT, allowing us to take into account origin passivation effects. The compatibility of experimental ORC and MCM data (i.e. potential origins) with optimized PODLS is variable, and suggesting to relax the hypothesis that the firing probability of potential origins is constant. Our approach therefore supports a combined origin density and affinity model and provides a basis to investigate the distinct genetic and epigenetic determinants of origin licensing and firing.

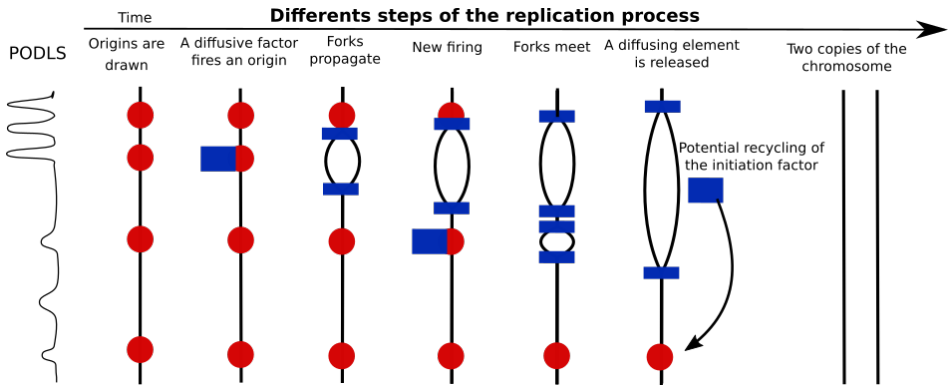

Fig 1. Modeling DNA replication. Given a PODLS possibly derived from a specific genomic feature (e.g. a DNase I HSS profile), a fixed number of localized potential origins is drawn (red circles). Limiting firing factors (blue rectangles) activate origins in a probabilistic manner and engage with each 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 replication a termination event and 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.

Results

Information complementarity between MRT and RFD profiles

Previous works on human DNA replication modeling [19, 34] compared simulated and experimental MRT profiles to substantiate the proposed models and constrain their parameter values. RFD profiles have now been established for many human cell lines [45, 59], providing us with an alternative comparison point. It is thus of interest to compare the information content of RFD and MRT profiles. Within the hypothesis of a constant fork speed v, MRT and RFD profiles are equivalent as they are deterministically related to one another by [22, 4, 2]:

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

Here $MRT(x)$ is the mean replication time after entry in S-phase of bin x and is expressed in time unit. Note that $MRT(x)$ as measured by Repli-seq experiments is the average replicated fraction at the moments locus x is replicated, and thus has a value between 0 and 1 that is converted into time unit by multiplication by an estimation of S-phase duration T_S. Equation (1) can be checked using experimental data. However, the derivative of a noisy profile is ill-defined and numerically unstable to compute. To avoid computing the MRT derivative, Eq. (1) can be integrated at point x over a length l leading to:

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

where $\langle\cdot\rangle_x^{x+l}$ stands for the average value over $[x, x+l]$. Equation (2) predicts that the MRT change across an interval is proportional to the average RFD over that interval. Using

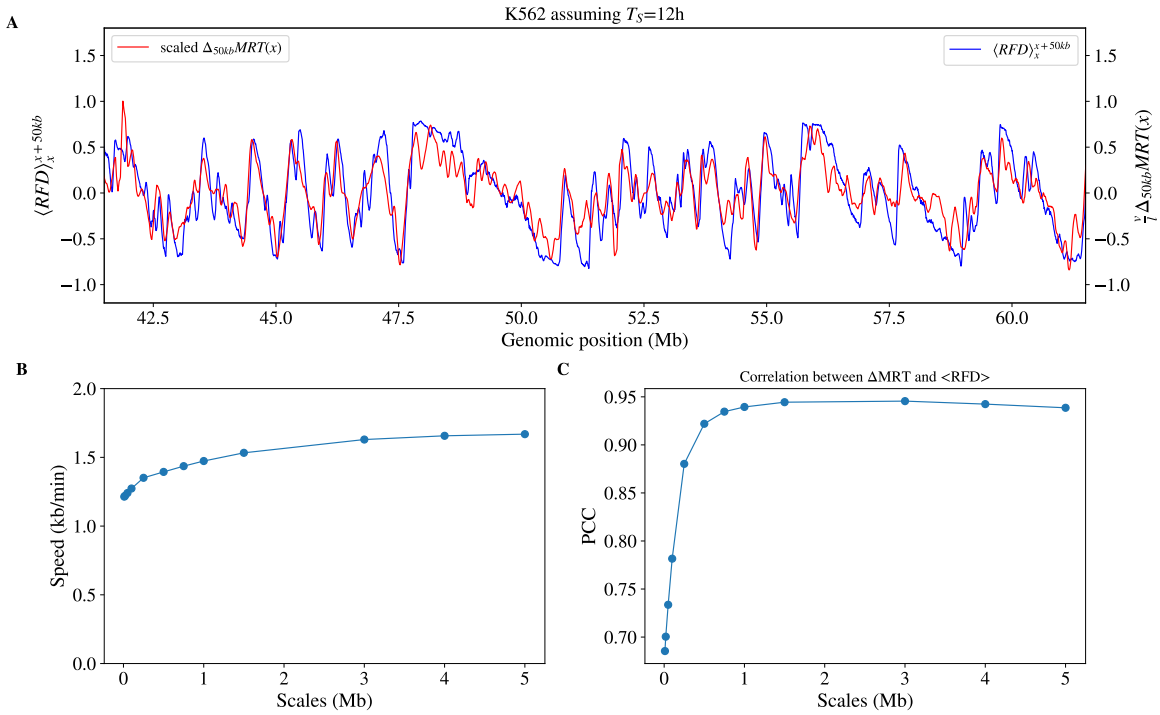

Fig 2. (A) Comparison, for a 20 Mb region of chromosome 1, of the K562 RFD profile averaged over 50 kb windows (blue; $\langle RFD \rangle_x^{x+50kb}$) with K562 MRT changes across 50 kb intervals ($\Delta_{50kb}MRT(x)$), following Eq. (2) with $v = 1.24$ kb.min^{-1} assuming $T_S = 12$ hour. (B) Replication 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.

reported Repli-seq MRT [25] and OK-seq RFD [59] profiles for the K562 cell line, Eq. (2) was very convincingly verified over a large scale range from 10 kb to 5 Mb, with a genome-wide correlation coefficient up to 0.94 at scale 1 Mb, and a proportionality coefficient ranging from $v = 1.2$ kb.min^{-1} to $v = 1.6$ kb.min^{-1}, assuming $T_S = 12$ hours [56]. This is illustrated on Fig. 2 for scale 50 kb and Fig. S1 for other scales. Therefore, although OK-seq and Repli-seq experiments are complex and have been performed by different laboratories, they are highly consistent with each other, on a wide range of scales, within the hypothesis of a constant fork speed.

In their modeling work, Gindin et al [19] found that of all epigenetic features tested, IPLSs based on DNase I HSS profiles produced the best match between simulated and experimental MRT profiles (Pearson correlation coefficient, PCC = 0.865). We performed similar simulations, using the related model outlined in the Introduction (Fig. 1) and detailed in Materials and Methods. Using a PODLS based on the K562 DNase I HSS profile, we drew a fixed number of potential origins and simulated a bimolecular reaction with firing factors, whose number increased from the start of S phase and plateaued after ≈ 1 h. Productive initiation events trap the factors at the forks and termination events released them, making them available for new initiation events. After grid search optimisation of the number of potential origins and the number of firing factors, we observed a high correlation (PCC = 0.88), similar to Gindin et al ([19]; 0.865), between simulated and experimental MRT profiles,

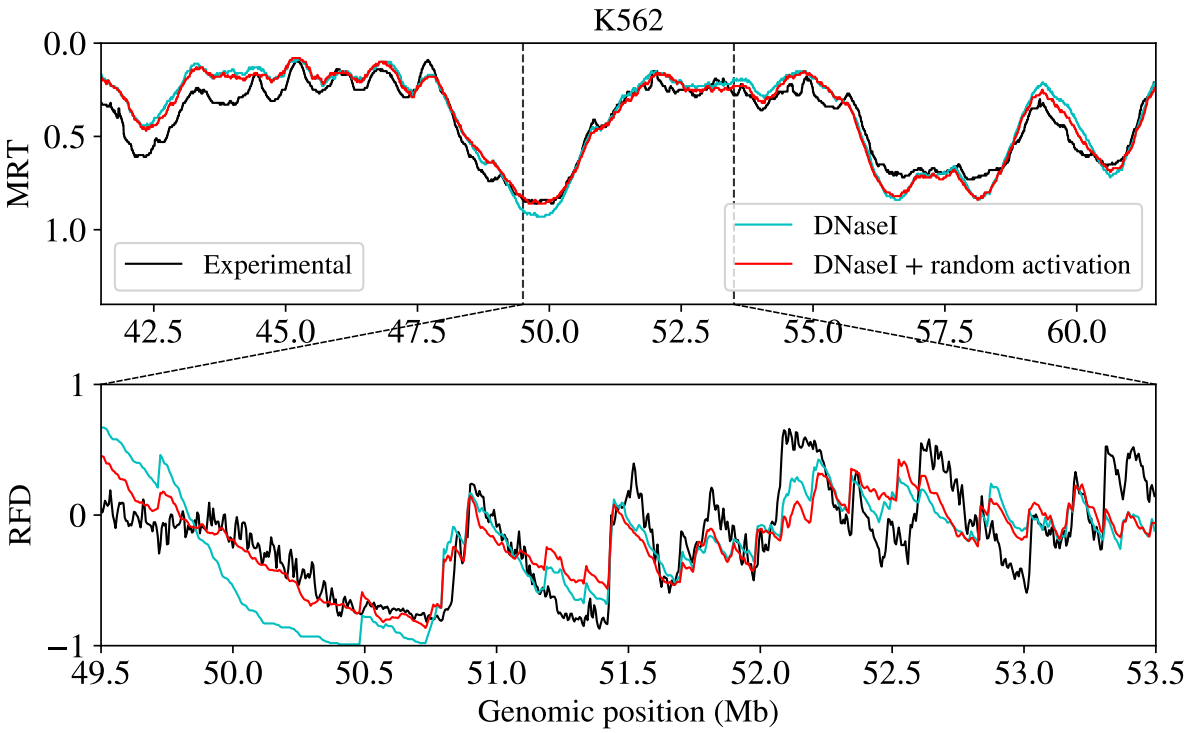

Fig 3. Comparison of experimental (black) and simulated (red, light blue) MRT profiles for a $\approx$ 20 Mb region of chromosome 1 (top) and RFD profiles for a 4 Mb region centered in the middle of the 20 Mb region (bottom) using a PODLS based on DNAse I HSS, with (red) or without (light blue) the addition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

and a lower correlation between simulated and experimental RFD profiles (PCC = 0.70) (Fig. 3). Reasoning that addition of dispersive, random initiation to the PODLS (see Methods) might improve the results, we extended the grid search for this parameter and obtained an optimal correlation for RFD at PCC = 0.75 for 5% of random initiation events, while maintaining the same correlation with MRT (PCC = 0.88) (Fig. 3). These observations confirm that MRT and RFD data are consistent with each other and suggest that RFD data are furthermore informative about random initiation.

Despite the theoretical equivalence of MRT and RFD profiles (Eqs. (1) and (2)), their correlation (Fig. 2) decreased at small scales, due to the low (100 kb) resolution of MRT profiles and, to a lower extent, at large scales, because integrating RFD sums up its experimental noise. In fact RFD profiles provides better information on origin positions, while MRT better reflects integrated origin activity over broad regions. This is illustrated by the following numerical experiments. When the positions of DNase I HSS were resampled within each 200 kb window prior to constructing the PODLS, the simulated MRT profile retained a high correlation with the experimental MRT (PCC = 0.87; Fig. 4A, green curve), while the correlation between simulated and experimental RFD profiles dropped (PCC = 0.61; Fig. 4B, green curve). The exact positions of DNase I HSS were critical to reproduce RFD profiles upward jump positions, which is consistent with the observed enrichment of OK-seq IZs with DNase I HSS [45]. On the other hand, the tolerance of MRT profiles to DNase I HSS resampling suggested that MRT is not sensitive to the precise origin positions within a

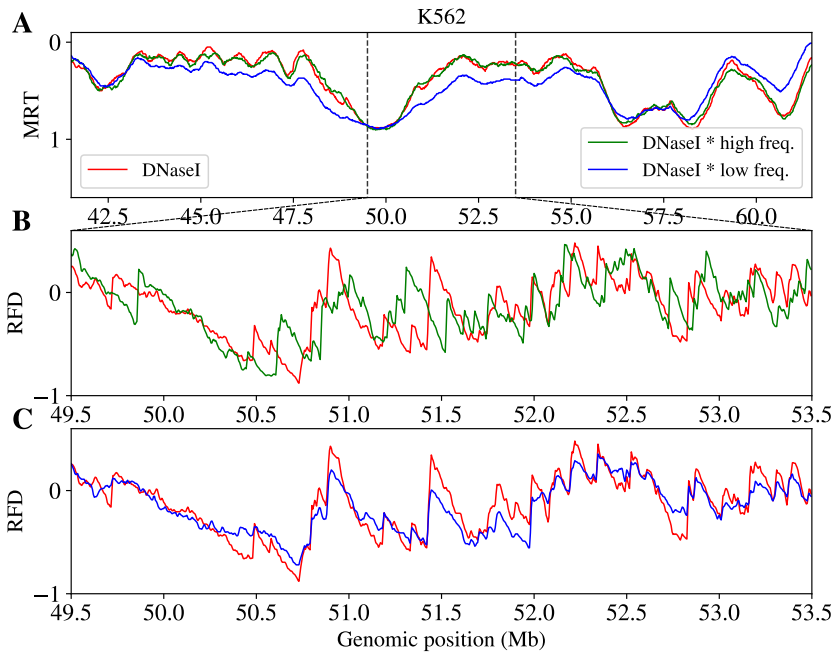

Fig 4. Comparison of simulated MRT (A) and RFD (B,C) profiles corresponding to different PODLS profiles all other model parameters being kept constant. PODLS were derived from (i) experimental DNase I HSS data (red), (ii) the DNase I HSS data after random shuffling of HS sites position within all 200 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).

sampling window.

On the other hand, although MRT can be computed by integrating RFD (Eq. (2)), this cumulates the experimental noise, blurring large scale features that MRT data more directly capture. The lesser sensitivity of RFD than MRT to large scale patterns was revealed in a second numerical experiment where we modulated the DNase I HSS signal amplitude with a slow varying function of large period ($P = 25$ Mb) before constructing the PODLS. In that setting, the correlation between simulated and experimental profiles decreased markedly for MRT (PCC = 0.72) but only slightly for RFD (PCC = 0.72) (Fig. 4AC, blue curves). Therefore, MRT is constrained by the collective action of multiple origins, so that the sum of neighbouring DNase I HSS signals is critical, while their exact positions per 200 kb windows are not (Fig. 4A). RFD is instead sensitive to the location rather than the magnitude of these signals. The influence of a single origin on RFD decreases rapidly with distance, due to the influence of intervening origins.

To summarize, incorporating RFD as a target for simulations likely allows to test origin positioning at much higher resolution than was achieved with MRT [19, 34]. Deriving RFD profile from MRT data (Eq. (1)) by numerical derivative would produce low resolution RFD

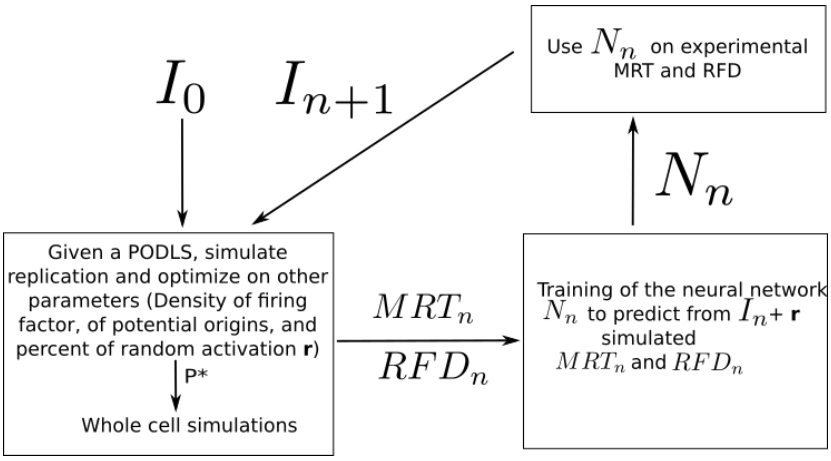

Fig 5. Schema of the iterative optimization procedure of the PODLS for simultaneous prediction of experimental MRT and RFD data. The starting PODLS I_0 may be a crude approximation of the target PODLS such as given by the peaks of RFD derivative, or the DNAse HSS profile, but not a random profile. We 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).

profiles with amplified noise, while determining MRT profile from the summation of RFD data (Eq. (2)) would produce MRT profiles with unreliable MRT changes over large distances. Experimental MRT and RFD profiles thus provide complementary information. We use both in the following analyses.

Learning the PODLS to optimize the prediction of both MRT and RFD experimental data. Having shown that experimental MRT and RFD profiles are consistent over a wide range of scales in the context of the constant fork speed hypothesis, we assessed to which extent they could be jointly explained by a single PODLS in our replication model. At 5 kb resolution, the PODLS correspond to ~ 575000 parameters which must be optimised. To achieve this, we designed an iterative method that progressively improves the PODLS (Fig. 5). It uses model simulations to train a neural network to predict the PODLS given the MRT and RFD profiles, i.e., to invert our replication model. We initialise the method by setting non zero values of the PODLS in regions with the highest RFD derivative (See Materials and Methods) i.e. in the strongly ascending segments of the RFD profile corresponding to the main initiation zones previously describes [45]. This crude approximation of the PODLS is named I_0. Then a grid search optimisation on the other parameters $P = (\rho_F, d_{PO}, r)$ of the replication model (See Material and methods) was performed. To limit computation time, this optimisation is performed over chromosome 2 only and results in a set of optimal parameters P_0 that maximized the sum of the Pearson correlation coefficients between experimental and simulated MRT and RFD profiles. Then we simulated whole genome replication using I_0 and P_0 to generate MRT_0 and RFD_0 and trained a neural network (See Materials and Methods) to predict $I_0 + r_0$ from MRT_0 and RFD_0, where r_0 is the optimal fraction of random initiation events given I_0. We then used this network to predict I_1 from experimental MRT and RFD, reasoning that I_1 should produce more accurate prediction of MRT and RFD than I_0. Another grid search optimisation on P, given I_1, was

performed to obtain P_1 and simulate MRT_1 and RFD_1, then train a new neural network to predict $I_1 + r_1$ and in turn I_2 form experimental MRT and RFD. These steps were iterated four times, because we observed that the correlations between experimental MRT and RFD profiles and their simulated estimates never improved with further iterations.

We first applied the procedure to K562 data, the sequences of joint correlations between experimental MRT and RFD and simulated profiles $(MRT_0, \cdot, MRT_4)$ and $(RFD_0, \cdot, RFD_4)$ were (0.79, 0.95 ,0.98 ,0.98, 0.98) and (0.8, 0.90, 0.91, 0.91, 0.91), respectively. We ran a grid search on the whole genome given I_4, this yielded unchanged correlation for both MRT and RFD, suggesting that parameter optimization on chromosme 2 only was not a limitation. We also tried using K562 DNase I HSS as the I_0 of the method. The I_0 profiles obtained from DNase I HSS or RFD derivative peaks presented some differences (PCC=0.76), but lead to very similar I_4 profiles (PCC=0.94; Supplementary Figure S2) and produced identical high correlations between simulated and experimental MRT (0.98) and RFD (0.91) profiles. In contrast, we were unable to ameliorate the PODLS starting from a random I_0 (MRT and RFD correlations were 0.67 and 0.84, respectively, using I_2, but decreased at step 3). Therefore, our optimization method required some initial information about the PODLS, but converged to nearly the same optimized PODLS from heterogeneous starting points. This is not a constraint as an adequate initialisation can be obtained from experimental RFD data.

To test the robustness of this inversion procedure, it was applied to replication profiles for GM and HeLa human cell lines and yeast *Saccharomyces cerevisiae*. It systematically resulted in high PCC between experimental and simulated profiles after the fourth iteration: for MRT 0.98 with GM; 0.98 with HeLa and 0.96 with *S. cerevisiae*; for RFD 0.92 with GM ; 0.85 with HeLa RFD and 0.91 with *S. cerevisiae* (see Supplementary Table S1 for the results of the different iterations). Figure 6 visually illustrates the striking consistency between simulation and experiments for the three different cell lines, not only in a correlative manner but also quantitatively. Correlation for HeLa RFD profile was less than other cell lines. Indeed HeLa is a priori a more challenging case as it has about twice as many IZs as K562 and GM [45, 59], but we also observed that some regions of poor RFD prediction corresponded to regions where experimental MRT and RFD profiles were incoherent possibly due to the difference between Hela cell lines used in different laboratory (Fig. 6).

Sensitivity of MRT and RFD with respect to model parameters.

We performed simulations of the replication of the whole genome to assess the sensitivity of MRT and RFD profiles as well as S-phase duration with respect to: (i) the density of firing factors ρ_F (number per Mb); (ii) the mean distance between potential origins d_{PO}; (iii) the proportion of random initiation r; and (iv) fork speed v. Working around the reference set of parameters ($\rho_F = \rho_F^* = 0.55$ Mb^{-1}; $r = r^* = 0\%$, $d_{PO} = 5$ kb, $v = 1.5$ kb.min^{-1}), we let one of the parameter vary (Fig. 7). ρ_F^* and r^* are the optimal values obtained at the end of the iterative procedure, without the final grid search exploration on the whole genome as it did not improved the correlations; the values for d_{PO} and v are reasonable choices (see discussion below).

The correlations of simulated with experimental MRT and RFD profiles showed a clear maximum at $\rho_F = \rho_F^* = 0.55$ Mb^{-1}, more pronounced for RFD (Fig. 7A,B). This number implied that the maximum density of forks was ~ 1 fork per Mb, or $\sim 6,000$ forks per diploid nucleus, at any time in S phase. This was in reasonable agreement with upper estimates of the maximal density of limiting factor Cdc45 (3.5 molecules per Mb) in mammalian cells [58],

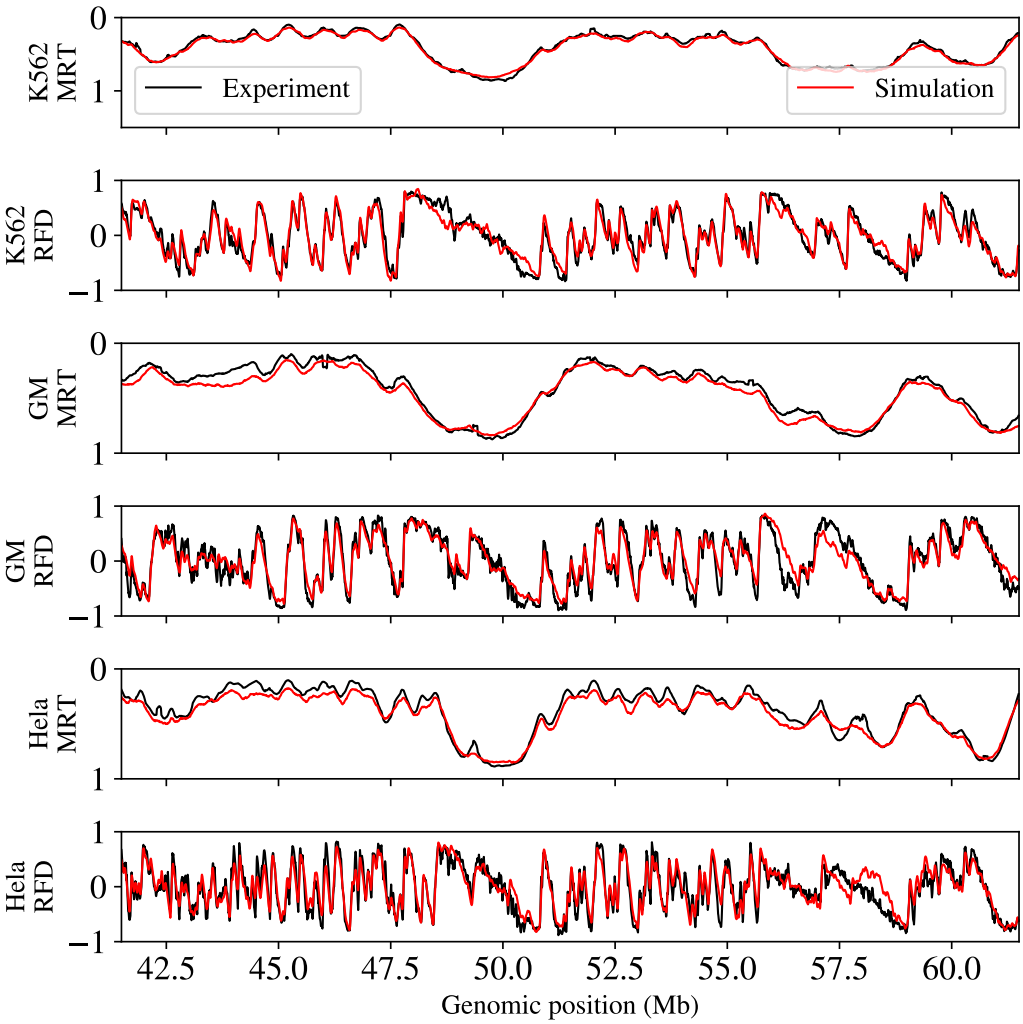

Fig 6. Comparison for the three indicated cell lines of experimental MRT and RFD (black) and simulated MRT_4 and RFD_4 profiles (red). A representative 20 Mb segment of chromosome 1 is shown. Note that the HeLa region of strong discrepancy between experimental and simulated profiles (between 57.7 and 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.

considering that 2 molecules of Cdc45 molecules are required to activate two diverging forks, one limiting factor in our model thus corresponding to a pair of Cdc45.

The optimal value for the random initiation $r^* = 0\%$ is confirmed as increasing r up to 20% slightly decreased both MRT and RFD correlations (Fig. 7G,H). The value $r = 0\%$ means that random initiation was correctly learned by the iterative method and did not require further external addition. In order to further apprehend the requirement of a minimal amount of initiation in all location, we performed the following experiments: (i) the random initiation was treated as an external parameter in the iterative PODLS optimisation procedure i.e., it was excluded from the training, the neural network was trained to output I_n instead of $I_n + r$; in that case the correlation with MRT and RFD was maximal for 5% of random initiation

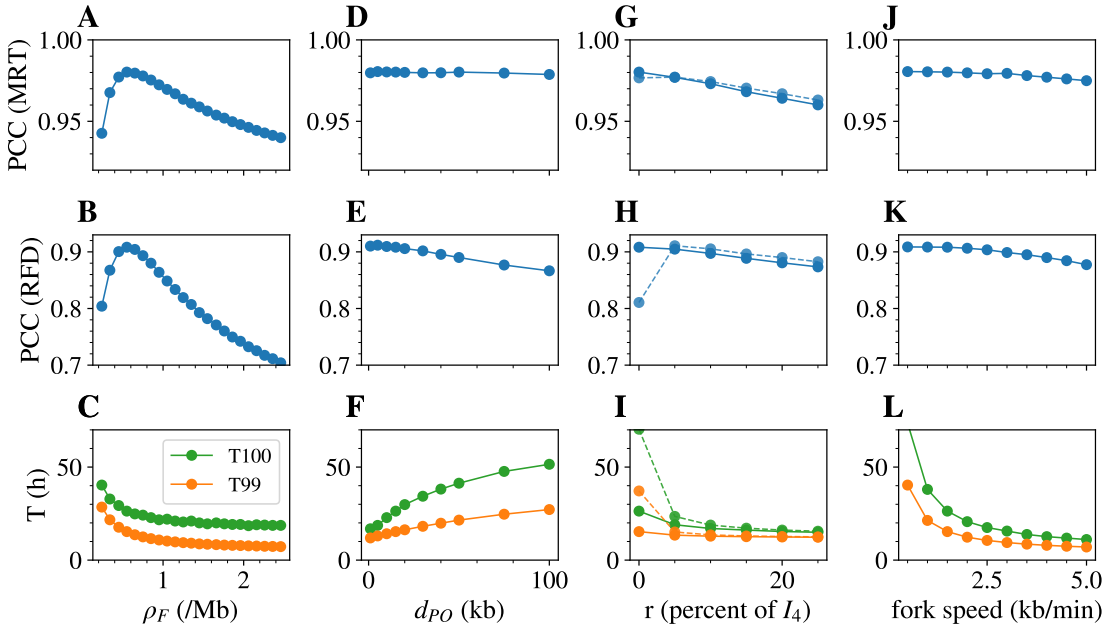

Fig 7. Effect of single parameter variation on measurable targets in K562, other parameters being kept as their value in the reference parameter set. Effect of the density of firing factors ρ_F (A,B,C); the average distance between potential origins d_{PO} (D,E,F) ; the proportion of random initiation 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 T99 (orange) and T100 (green), the times required to replicate 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.

(Fig. 7G,H); (ii) setting to zero the lowest I_4 bin values totalling 5% of potential origins ($\approx 53\%$ of the bins), significantly decreased the correlation with RFD data to PCC $= 0.82$: without these random initiations, the simulations fail to capture the RFD in large late replicating regions whom replication time is also clearly delayed (Supplementary Figure S5). The null RFD profiles in these regions are indeed compatible with random initiation.

Replication fork speed in human cell lines is reported in the range 1-3 kb per min and the average inter potential origins distance is expected significantly smaller than 50 kb as the average distance between activated origin has been reported in the 50-100 kb range and potential origins are likely in large excess of the actually fired origins. Variations of d_{PO} and v over these reasonable ranges of values ($v \in [1, 3]$ kb min^{-1}, $d_{PO} \lesssim 15$ kb) had practically no incidence on the correlation of experimental and simulated MRT and RFD; for larger values of these parameters a small decrease of the PCCs is however observed (Fig. 7D,E,J,K). Thus these parameters' values are only very weakly constrained by the experimental MRT and RFD profiles, which justify why they were left from the parameter optimization in the iterative PODLS optimization procedure.

These results were robustly observed when a similar analysis was performed for GM and Hela replication data, the position of the maximum values for ρ_F^* where of 0.45 and 0.82

Mb^{-1}, respectively (Supplementary Figures S6 and S7). In summary, MRT and RFD data, being highly consistent between themselves, are jointly and precisely explained by a simple model featuring a unique PODLS input and values for d_{PO}, v and ρ_F in agreement with the current knowledge. A small amount $\sim 5\%$ of random initiation was necessary to account for experimental data -i.e., any region of the genome has a minimal replication initiation potential.

Dependence of replication kinetics with respect to model parameters.

We first analyzed K562 S-phase duration T_S using the median of the times to replicate 99% and 100% of the genome, T99 and T100 respectively. As expected, T_S decreased with the density of firing factors ρ_F and the fork speed v (Fig. 7C,L) as we are in a regime of strong affinity between firing factors and potential origins (large k_{on}) so that $T_S \approx \frac{1}{2*v*\rho_F}$ [1]. The model predicted important differences between T100 and T99, as the latest replicated regions were the ones most devoid of potential origins. The predicted distance between potential origins as a function of MRT significantly increases from early to late MRT: for a genome-averaged distance $d_{PO} = 15$ kb, the average distance in very early regions (MRT < 0.1) was as short as 2 kb compared to 216 kb in very late (MRT > 0.9) regions (Supplementary Figure S3). This observation also explained that (i) the cell-to-cell variability of T99 (~ 10 min) were much smaller than for T100 (hours) (Supplementary Figure S4); (ii) increasing d_{PO} increased T_S to greater extend when considering T100 than T99 (Fig. 7F); and (iii) the addition of random initiation decreased T_S, only in a limited manner when considering T99, but with a strong effect on T100 which decreased abruptly as r was increased to 5%, and more slowly thereafter (Fig. 7I). When considering the PODLS optimized considering random initiation as an outside parameter, the latter effect is strongly enhanced as in this case the latest replicated regions are devoid of potential origins unless random initiation is added (Fig. 7I). These observations are reminiscent of common chromosomal fragile sites (CFSs), which replicate very late and show a paucity in origins that explains their fragility [33].

We used T99 as a more robust proxy of experimentally measured S phase lengths than T100, because (i) experimental detection of S phase markers may miss the earliest or latest S phase cells; (ii) very late replication of specific sequences, amounting to $\sim 1\%$ of total DNA, does occur during what is termed the G2 phase of the cell cycle [57]; (iii) within the parameter range we explore, T99 variations (10 min) were smaller than for T100 (hours) (Fig. S4). Experimental measurements of the cell-to-cell variability of S phase length also ranged from minutes to hours depending on detection methods [43, 55]. Using the reference set of parameters ($\rho_F = \rho_F^*$; $r = r^* = 0\%$, $d_{PO} = 5$ kb, $v = 1.5$ kb.min^{-1}) , the predicted T99 for HeLa and K562 were 9.4 h and 13 h, respectively, very close to the reported S phase lengths of 8.8 h in HeLa [23] and 12 h in K562 [56]. Taking account that a fork speed of $v = 2.0$ kb.min^{-1} has been measured in the JEFF cell line [51], closely related to GM06998, we obtained a T99 of 11.3 h for GM, consistent with an experimental estimate of 10 h [22]. In our simulations, the faster replication of HeLa cells was explained by a larger number of firing factors than in GM (2-fold) and K562 (1.5-fold). Thus, the optimisation procedure selected a density of firing factors ρ_F that gave S phase duration coherent with experimental measurements using sensible values for d_{PO} and v.

The origin firing rate per length of unreplicated DNA, $I(t)$, was previously reported to follow a universal bell-shaped curve in eukaryotes [1, 20] (Supplementary Figure S8). As

expected from the choice of the k_{on} value, the simulations produced the expected shape for the three cell lines with a maximum of $I(t)$, $I_{max} \simeq 0.15$ Mb^{-1}.min^{-1}, in reasonable agreement with an estimate of 0.3 Mb^{-1}.min^{-1} from low-resolution MRT data [20]. Finally, we measured in K562 the dispersion of replication times (RT) of each loci as a function of their MRT. A recent high-resolution Repli-Seq study [61] reported that RT variability estimated at each loci as the width between the first and the third quartiles of the locus RT distribution, increased from early to mid S phase and decreased from mid to late S phase. For most regions the RT variability was in the 1.25–2.5 h range. Our simulations in K562 produced a similar behavior of locus RT variability but over a wider range, from ~ 0 h in early or late S phase to 3 h in mid S phase (Supplementary Figure S10). In our model where non uniform MRT arises from non uniform PODLS, the replication time variability of a given locus is a consequence of the stochasticity of the model where potential origins positioned randomly according to the PODLS may fire at any time.

In summary, the kinetic parameters of S phase predicted by our stochastic model are (i) consistent with the shape of the dependency on time of the firing rate $I(t)$ and replication time variability and (ii) predictive of the S phase duration but RT variablity is higher than reported, suggesting that *in vivo* replication kinetics is less stochastic than our model.

Direct estimation of the PODLS from experimental data

Origin firing can be prevented by context-dependent passivation from nearby origins. An important concept in DNA replication modeling is therefore the distinction between observed origin efficiency (OE) and intrinsic origin efficiency (IE). OE is the fraction of origin copies that fire in a population, whereas IE is the efficiency that would be observed, in the absence of passivation, over the entire length of S phase.

In our model, potential origins are MCM DHs which all have the same elementary probability of firing. For one bin x with $n(x)$ MCM DHs, given the reaction rate k_{on} and the number of free firing factor $F_{free}(t)$, the probability for firing to take place during an elementary time dt is:

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

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

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

Hence, with an infinite time to fire, $A_x(t)$ converges to one unless the locus is devoid of any potential origins ($n(x) = 0$). The observed firing efficiency OE is smaller than one due to passivation. Reasoning that replication of a locus at a time earlier than average ($RT < MRT$) typically occurs when an origin is activated in the locus and that $RT > MRT$ occurs when a locus is passivated, we expect $OE(x) = A_x(MRT(x))$ which leads to:

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

We first assessed the validity of this relationship in the set of 200 S-phase simulations using the optimized PODLS I_4 and the reference set of model parameters in K562. $OE(x)$ and $MRT(x)$ were the averages of the recorded activation status and RT in each simulations. The

total number of MCM DHs over the genome being L/d_{PO}, the average MCM DH density profile $n(x) = I_4(x) * L/d_{PO}$ (MCM DH/5kb). Finally we used $[F_{free}] = 40$ the observed almost constant value for $MRT < 0.7$ (Supplementary Figure S11) in K562 simulations. The two terms of Eq. (5) were computed and we obtained a genome-wide Pearson correlation coefficient of 0.71 and a proportionality coefficient of 0.68. Equation (5) works reasonably well but tends to underestimate observed origin efficiency. Human IZs were delineated as the region with significantly increasing RFD profiles [45], thus corresponding to regions where ΔRFD is locally maximal. When focusing on the local maxima of ΔRFD ($\sim$ 9800 peaks in K562, Materials and methods), we observed that the predictive power of Eq. (5) was better than for the whole genome: the PCC improved to 0.81 and the coefficient of proportionality to 0.87 between the two terms of Eq. (5) (Supplementary Figure S13). We also computed simulated Repli-seq MRT as can be derived from Repli-seq experiments multiplied by $T_S = T99 = 13$h and observed that it was a very good approximation of true MRT (Supplementary Figure S13 so that the correlation coefficient as well as proportionality coefficient between the two terms of Eq. (5) where not noticeably change by using Repli-seq MRT instead of MRT (Supplementary Figure S12). These results indicate that our hypothesis that $OE(x) = A_x(MRT(x))$ applies more precisely at IZs determined as ΔRFD local maxima and that MRT can be substituted by Repli-seq MRT and T_S.

Generally speaking, the ΔRFD across a genomic segment is twice the difference between the number of initiation and termination events in the segment [2]. Mammalian IZs are broad and may contain multiple MCM DHs. Termination events are nevertheless rare or absent within IZs [45, 24]. Hence, one can estimate $OE(x) = \Delta RFD(x)/2$ in these loci. For example, at a locus with a 100% efficient origin, the RFD will shift from -1 to +1, creating a jump of $\Delta RFD = 2$. This was indeed observed in our K562 simulations when focusing on the local maxima of ΔRFD ($\sim$ 9800 peaks in K562, Materials and methods): PCC was 0.97 with a coefficient of proportionality of 0.9 between $OE(x)$ and $\Delta RFD(x)/2$ (Fig S15). As a consequence, Eq. (5) linking $OE(x)$, $n(x)$ and $MRT(x)$ provides a link between $RFD(x)$, $n(x)$ and $MRT(x)$ in IZs suggesting a simple and direct way to estimate the PODLS ($\propto n(x)$) from MRT and RFD profiles:

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

where $n_e(x)$ is the $n(x)$ profile estimated from the measurable parameters ΔRFD and MRT up to a multiplicative constant. Note that for small value $\delta RFD(x)$ (for 5 kb bins, the maximum value of $\Delta RFD/2$ is 0.16), one can use a Taylor extension of ln to simplify Eq. (6) yielding:

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

We then compared $n_e(x)$ estimated using simulated MRT and RFD with $n(x)$, the input of the simulation (Fig. 8 A). The Pearson correlation was 0.94 and the coefficient of proportionality was 0.82 at the 9,800 ΔRFD local maxima. We observed from further exploration of the dependence on $MRT(x)$ in Eq. (6) (Fig. 8 B) that behavior of $n(x)$ in late MRT was better capture by an exponential dependence rather than the inverse linear dependency on MRT of Eqs. (**??**), leading to the empirical relationship:

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

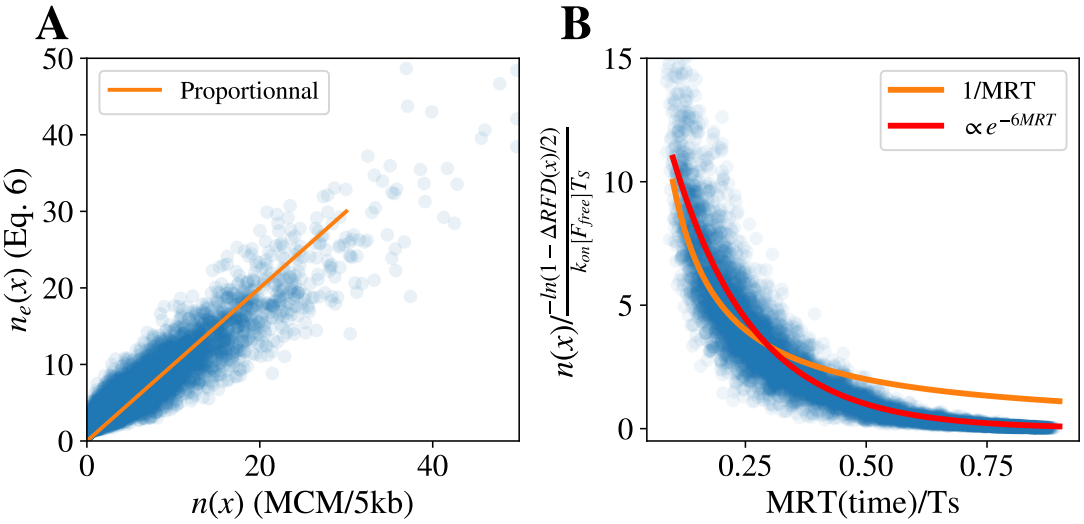

Fig 8. (A) Comparison of the average number $n(x)$ of MCM per 5kb bin used in the simulations, and the predicted number of MCM $n_e(x)$ using Eq. (6) at the 9800 ΔRFD peaks in K562. (B) Scaled $n(x) \times k_{on}[F_{free}]T_S/(-\ln(1-\Delta RFD(x)/2)$ as a function of Repli-seq MRT at the 9800 ΔRFD peaks (blue circles) in K562 simulations; it is compared with $1/MRT$ (orange) and proportionality to exp^{-6MRT} (red).

better captured this late dependency (PCC=0.96). In Eq. (8) the MRT dependency was even stronger than in Eq. (6). We hypothesise that Eq. (8) accounts for the actual increase of $F_{free}(t)$ at the end of S-phase (Supplementary Figure S11 that boosts potential origin firing efficiency so that a smaller number of MCM are required to produce the same OEs.

We then wondered if it would be possible to simulate replication using a PODLS directly derived from experimental data using Eqs. (6) or (8) (not that since PODLS is normalized the ignorance on the prefactors is not an issue). For this we selected the 15% highest experimental $\Delta RFD(x)$ (Materials and Methods) and used (6) to predict the PODLS that we used in our model. The resulting simulated MRT and RFD profiles were highly correlated with experimental profiles (Table 1), for example PCC=0.94 between MRT and 0.88 between RFD profiles in K562. Interestingly, this formula for PODLS prediction robustly applied also in *S. cerevisae* with Pearson correlations of 0.93 for MRT and 0.90 for RFD. PODLS derived from Eq. (8) led to even higher correlations between simulated and experimental MRT (PCC=0.97) and RFD (PCC=0.91), very close to the correlation coeffecients obtained using I_4 profiles (Table 1). We also confirmed this result in *S. cerevisae* with PCC=0.96 for MRT and PCC=0.9 for RFD. These results show that measuring local RFD upshifts as OE estimates and MRT is sufficient to produce a near optimal PODLS.

SNS-seq or Bubble-seq signals provide in principle a direct estimate of OEs profile up to a proportionality coefficient that can be used in the same way as OE estimates in IZs from ΔRFD (above). Using such data (SNS-seq K562, SNS-seq Hela and Bubble-seq GM06998) directly as PODLS resulted in poor correlations between simulated and experimental MRT and RFD profiles (MRT, PCC=0.54 , 0.43 and 0.30 and RFD, PCC=0.16 , 0.23 and 0.12, respectively). Combining with the MRT information to infer the PODLS using Eq. (8), resulted in improved correlations (MRT, 0.81 , 0.87 and 0.83 and RFD 0.48 , 0.59 and 0.53,

respectively) (Table 1 1). However, the PODLS derived using the same MRT information from a flat OE profile resulted in better correlations than SNS and Bubble profiles (MRT, 0.97 , 0.98 and 0.98 and RFD 0.71 , 0.62 and 0.71, respectively), suggesting that these data do not improve on MRT profile information to the same extent as RFD profiles do.

We also analysed experimental data directly predicting potential origin positioning (Table 1). Using PODLS computed from K562 ORC2 [39], HeLa MCM7 [49] and HeLa MCM2 [17] resulted in Pearson correlation coefficients between simulated and experimental profiles of 0.87, 0.46 and 0.54 for MRT and 0.74, 0.28 and 0.43 for RFD, respectively. Finally, we computed the correlation of the SNS-seq, Bubble-seq and ORC/MCM signals with the optimize PODLS I_4 in the corresponding cell line at 5 kb and 50 kb resolution. For K562 SNS, Hela SNS and GM Bubbles, the correlations at 5 kb and 50 kb were 0.06 and 0.16, 0.18 and 0.31, and 0.06 and 0.17, respectively. For K562 ORC2, Hela MCM7 and HeLa MCM2, the correlations were 0.29 and 0.53, 0.18 and 0.31, and 0.27 0.41, at 5 kb and 50 kb resolution, respectively. Therefore, none of these experimental datasets were convincing predictors as RFD upshifts and MRT (Eq. (8)) of the PODLS I_4 coherent with experimental MRT and RFD profiles, for reasons that remain to be elucidated. The fact that ORC location seem to better predict the PODLS than MCM is somewhat unexpected, given experiments in Xenopus egg extracts that suggest that the multiple MCM DHs loaded around a single ORC binding site are equally competent to bind Cdc45 and trigger origin firing [26, 15].

The observed discrepancies between predicted and experimental potential origin density profiles may simply be explained by spatial non-uniformity e.g., locus dependent k_{on} or temporal non uniformity e.g., time dependent F_{free}. In the first case, the position-dependent landscape of k_{on} would be given by the I_4/MCM predicted over experimental potential origin density ratio. In the second case, this ratio should vary with time but not with genomic location at constant MRT. We found that the normalized I_4/MCM ratio computed in 50 kb windows (excluding null MCM windows; Fig. S16) decreased with MRT. Therefore, MCM DH activation was easier in early compared to late S phase, which could be partly explained if the number of firing factors was higher in early S phase than in our simulations and decreased thereafter. In addition, we observed a broad dispersion around this trend, suggesting that even at similar MRT, all MCM DHs are not equally reactive to firing factors. This could be due to heterogeneities in the MCM DHs themselves or in their surrounding chromatin environment, or to experimental noise.

Experimentally, the maximum RFD upshift observed at a 5kb resolution is of 0.16, corresponding to a domain of 8% efficiency. According to Eq. (7) and k_{on} and F_{free} from our simulations, at $d_{PO} = 5$ kb, this would correspond to 3 MCM DHs per 5 kb in late replicating (MRT=11h) domains. According to Eq. (7), the same number of MCM DHs would be only 0.8% efficient if placed in an early-replicating domain (MRT = 1 h), due to faster passivation. To observe OE = 90% in early such early locus, it would need to increase its intrinsic efficiency by a factor of ~ 10, i.e. by loading 34 MCM DHs instead of 3, if k_{on} $[F_{free}]$ is kept constant. Given that MCM DHs occupy 60-62 bp each and are located in linker DNA between nucleosomes [17], a 5 kb chromatin segment cannot accommodate more than 25 MCM DHs, which is lower than 34. This steric constraint could be a reason why the maximum observed OE in 5 kb bind is 8% . As a consequence, in order to be 90% efficient, an early domain may thus be $\gtrsim 100$ kb long. While this is still compatible with the observed size range of efficient IZs [45], increasing k_{on} or $[F_{free}]$ in early S phase, as Fig. S16 suggests is the case, would relax this constraint.

To summarize, we checked that OEs can be accurately measured from RFD upshifts and

PODLS	K562 MRT	K562 RFD	GM MRT	GM RFD	Hela MRT	Hela RFD
I_4	0.98	0.91	0.98	0.91	0.98	0.85
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}	0.97	0.71	0.98	0.71	0.98	0.62
ORC2	0.87	0.75	–	–	–	–
SNS e^{-6MRT}	0.81	0.48	–	–	0.87	0.59
Bubble e^{-6MRT}	–	–	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.

we could predict from these OEs and MRT the number of potential origins (MCM DHs) per 5 kb bin in IZs assuming a value for d_{PO}, thanks to our model where all spatial heterogeneity results from lisencing as captured by the PODLS (all MCM DHs have the same locus- and time-independent probability of firing per unit time). The observed discrepancies between predicted and observed MCM DH densities support a mixed, potential origin density and affinity model (MCM DHs do not all have the same affinity for firing factors). Future investigations of the $I4/MCM$ simulated over experimental potential origin density ratio should help discriminate the licensing and post-licensing mechanisms that regulate origin firing probability.

Discussion

Information complementarity and coherency between MRT and RFD MRT and RFD profiles in a constant fork speed hypothesis account for the same information. Here we have shown that it is possible to compare delta of MRT with the integral of RFD with only one free parameter, the fork speed. The rather narrow range of values obtained by fitting the fork speed (from ≈ 1.2 to ≈ 1.6 kb/min) over the large range of scale explored (from 5kb to 5 Mb), and the high correlation obtained (from 0.68 to 0.95) suggest that the experiment are compatible, even when considered at resolutions lower that what was expected (MRT resolution was estimated to be 100kb) and that the hypothesis of a constant fork speed at least on the scales considered (higher than 5kb) is robust. We have also shown by randomly resampling HS DNaseI site that RFD profiles contain information at higher resolution than MRT profile. However MRT profiles contain information about the strength of a whole domain, information that is lost with RFD profiles due to integration of the noise of RFD profiles. These two observations led us to the use of both profiles for our inversion method.

Inverse method Different mathematical models have been developed to estimate origin intrinsic efficiency from MRT data [40] or Genome-Wide Replication Timing Data [3], and more recently from RFD [5]. The three models [40, 3, 5] assign either a discrete number of origins [40, 5], each having probability to be activated as a function of time, or a continuous spatio temporal initiation density [3]. In this type of models, the hypothesis are more complex than in our simple model. First the number of firing factor is freely adjusted along the S-phase, while in our case it is constant after a small time. A second hypothesis is that the probability to activate an origin change as a function of the time, and it thus more difficult to test the assumption that the replication efficiency is set at the end of licensing by modulating the number of potential origins. Furthermore, all of these methods are challenging to adapt to the human genome, because they require a non trivial optimisation that can be carried on Yeast Genome, or a Bayesian analysis [5] that can be carried on 3 origins at a time. Here we have trained a neural network on simulated replication profiles to estimate the PODLS from MRT and RFD. This approach is very fast on human genome and output a 1D profile of origin density, which allows to test if the replication strength can be controlled by the number of origin set at the end of licensing as shown in Yeast [11]. The numerical code for our method is available at ()

Random activation We found that the PODLS could be decomposed in peaks and a flat initiation that correspond to randomly localised origin as was needed in Miotto et al [39] . However the level of random activation was found much lower than from miotto [39] which in their analysis found 60% of random activation while in our case a random activation of the order of 5 % was necessary. This is coherent with a more recent report in Yeast S. Cerevisae [27].

PODLS The obtained PODLS allowed us to simulate replication profiles for the three different cell lines considered and the simulated MRT and RFD profiled where all very highly correlated to experimental ones, the highest correlation being 0.98 for the three MRT profiles and the lowest one 0.86 for RFD profile of Hela cell line. This show that this simple modeling framework with constant speed, almost constant number of firing factors and the stength of the replication set at the end of licensing is sufficient to account for most of the variability in these profiles. We then analysed these PODLS profiles and found i) that the number of MCM is proportional to the derivative of RFD ii) it is inversely proportional to the MRT, following either a $1/x$ dependency or $1/e^{6x}$ dependency. We derived analytically this formula and could thus directly extract the number of MCM factor from experimental MRT and RFD. We also showed that these laws hold in Yeast S. Cerevisae. This formula allows us to estimate that an IZ with a given δRFD in an early context, is 10 to 50 times stronger than the one needed to obtain the same RFD jump in a late context.

SNS-seq, Bubble-seq By exploring the realtion-ship beween IE and OE, we where able to show that using SNS-seq and Bubble-seq as IE in our simulation framework while they are OE, lead to low correlation with MRT and lower correlation with RFD ($\approx 0.1 - 0.25$). Correcting them to OE by using MRT data lead to high correlation with MRT, but also to higher correlation with RFD ($\approx 0.5 - 0.6$).

Establishing Replication efficiency by licensing an inhomogeneous number of origin Our model also support the fact that origin replication strength could be set at the end of licensing, coherently with what has been shown by Das et al [11] in Yeast, where the origin efficiency can be accounted for by the number of MCM loaded at each origin. While a constant but non uniform profiles is sufficient to recapitulated MRT and RFD profiles, in this modelling work it is not possible to distinguish between a non uniformity that would come from a non uniform distribution of MCM, or from a uniform distribution of MCM, but with a non uniform firing rate due to for example an inhomogeneous accessibility of the firing factors. In K562 the values of the correlation obtained using ORC2 which are about 0.1 lower in correlation than our best simulations suggest that the amount of ORC2 could contain all the information needed to account for replication inhomogeneity. Surprisingly however MCM profiles in Hela used as such, allowed to obtain an intermediate correlation with MRT and RFD profile (0.55 and 0.44 respectively). Extracting peaks of the profile allowed to reach an higher correlation of 0.77 and 0.6. However these correlations are about 0.2 lower than our optimum correlation and do not allow to conclude if MCM non uniform distribution alone account for replication inhomogeneity.

In our perspectives we plan to extend our analysis to other cell lines when the dataset will be available. The already obtained PODLS will allow to push forward the determination of epigenetic marks and mechanism responsible for replication inhomogeneity in further analysis.

Materials and Methods

Model and simulation

Model and parameters

We model the replication initiation process by a bimolecular reaction between free firing factors F and potential origins PO on the genome of size L with a reaction rate k_{on} . This rate is the probability per unit of time that a firing factor and a PO meet and than an initiation follows. Once an initiation event has occurred, two forks propagate in opposite direction at speed v and a firing factor is trapped. The number N_F of firing factor is fixed and parametrised as $N_F = \rho_F L$ with ρ_F the density of firing factors. A potential origin density landscape I is used to position $\frac{L}{/} d_{PO}$ origin prior to each S phase entry along the genome, with d_{PO} the genome average distance between potential origins. We also decompose the initiation probability landscape I in two terms: it is the sum of a profile I^e , that can be e.g, an epigenetic profile, and a uniform contribution which correspond to a percent r of the initial profile I_e. In all the simulations we fix $v = 1.5$ kb/min and $k^e_{on} = 8.625$ / min /kb, so that if we simulate the whole genome,(in our case, the first 22 human chromosomes whose total size is 2.875 Mb), then $k_{on} = 1e^{-6}$ min^{-1}

So a typical simulation has 4 free parameters (ρ_F,ρ_{PO},I_e,r).

Simulation implementation

The modeled genome is always considered at 5 kb resolution. The input spatial profile $I^e + r$ is normalised so that the sum is one. We draw from this normalised profile $N_{PO}(t = 0) = \frac{L}{d_{PO}}$ origins, with replacement, meaning that several origins can be drawn in the same 5 kb window. During the simulation we introduce $\rho_F L$ firing factor following an exponential characteristic law $\rho_F L(1 - e^{-t/\tau})$ with τ a characteristic time taken to 1 h to simulate

progressive activation of firing factor upon S-phase entry. We use a Gillespie algorithm [18] to simulate an initiation reaction with 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 algorithm consider that the the next reaction between an origin and a firing factor will take place after 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 to δt_E, the smallest time of encounter for two forks on the genome. The system then evolve for an increment of time $\delta t = min(\delta t_E, \delta t_A)$, meaning that all the forks on the genome moves from $v\delta t$. If $\delta t_A < \delta t_E$ then one origin is activated at random, a firirng factor is trapped ($N_F(t+\delta t) = N_F(t) - 1$), and two forks propagate on opposite directions at a velocity v from the origin. Otherwise the termination event releases a factor so that $N_F(t+\delta t) = N_F(t) + 1$ and a new time step begin. If a fork replicates a positioned, unfired origin, then this origin is passivated and remove from the list of potential origins.

Remark on rescaling the simulation for the simulation on chromosome 2 only. In [1] we showed that to reproduce the shape of the experimental temporal rate of DNA replication origin firing, one has to be in a regime governed by the critical parameter $\rho_F^* = \frac{v}{k_{on}Ld_{PO}}$. In order to stay in this regime no matter the size of the genome, we parametrise the simulation with $k_{on}^e = k_{on}L$ chosen constant. This means that if we increase the size of the system, the constant of reaction k_{on} must be decreased, which is coherent as k_{on} encompass the efficiency of encounter between one potential origin and one firing factor and being in a larger system their encounter rate is decreased.

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

Computing experimental quantities, comparison with experimental data. To compute RFD, we record in each simulation the fork direction as +1 -1 and each initiation or termination as 0. The final RFD is the mean over the 200 of simulations. To compute MRT, as the one in repli-seq experiment, we record for each simulation and each locus the actual replicated fraction of the genome at the time the locus is replicated. Then to simulate the six fractions of the repli-seq experiment, the continuous [0..1] interval of replicated fraction of the genome is converted into six regions of length 0.2 . Then $MRT(x) = \sum p_i(x)i/6 + 1/12$ where $p_i(x)$ if the fraction of the simulations where the locus at position x has been replicated when the replicated fraction of the genome was between $[i/6, (i+1)/6]$.

For both MRT and RFD when comparing with experimental data, we masked the region removed as specified in the missing data paragraph. Pearson correlations were computed at 5 kb resolution for RFD and 10 kb resolution for MRT. T_{100} was defined as the replication time of the latest replicated window T99 (resp. T95) we defined as the time at which 99% (resp. 95%) of the genome was replicated.

Experimental data

DNaseI HS experiment were downloaded from the ENCODE project [25, 52]: (K562 DNaseI HS Uniform Peaks from ENCODE/Analysis (narrowPeak))

ORC2: from [39] (supplementary material table S1 in the [39] article). ORC2 binding sites in K562 cells (Genomebuild: hg19)

SNS: K562 [46] (GSE46189_Ori-Peak.bed) , Hela: [7] (Hela_SNS_seq_Besnard_Tot.1kb.csv)

Bubble-Seq: [37] (GSE38809_GM_combined_RD_bubbles.bedgraph)

RFD data : Hela [45] , GM06990 and K562 from [59].

Repli-seq data: For mean replication timing data for GM12878, K562, HeLaS3, alignment files of Repli-seq libraries (BAM files) for six S-phase fractions were obtained from the ENCODE project [25, 52] at
http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwRepliSeq/.

Yeast data : OKSeq [28]. MRT was obtained from [42]. we shifted MRT by 0.05 for numerical stability.

Grid search optimisation

When performing grid search optimisation on the four parameters ρ_F, d_{PO}, r ,v we noticed, as explained in the main text, that both d_{PO} and v had little effect on the MRT and RFD profiles. These two parameters where thus left out of the grid search. Thus when a grid search optimisation is carried on it is done only on 2 dimensional grid ρ_F and r. The optimum selected is the one of highest sum of pearson correlation between simulated and experiment MRT and RFD. For the grid search optimisation carried on the paragraph (Learning the PODLS that best predicts both MRT and RFD data.) during the iterative procedure on the chromosome 2, for r, the explored values are [0,0.02,0.05,0.1], and for ρ_f the range is [0,6,0.8,1,1.2,1.4]*ρ_{start} where ρ_{start} is the density of firing factor needed to replicate the whole genome at a fork speed 1.5 kb.min^{-1} for a S-phase duration T_S if all the firing factor are active: $\rho_{start} = \frac{1}{2*v*T_S}$. $T_S = 8$ h for Hela cell and 12$\tilde{\text{h}}$ for GM and K562.

When performed genome-wide (Results reported in Tab 1), and to be able to compare all cells, we choose the same grid for all human cell lines: The percent of random activation vary from 0 to 20 % by increment of 5 %. The density of Firing factor was explore in the range [0.27 0.41 0.55 0.68 0.82 0.95 1.1] firing factor per Mb.

Iterative procedure used in Learning the PODLS that best predicts both MRT and RFD data.

Our iterative method aims to extract an initiation landscape coherent with both MRT and RFD profile given a model of replication. The first step is to extract a starting initiation profile. We computed the increment of RFD between 5 kb bins, then smoothed the profile using an average sliding window over 50 kb to remove noise and keep the Values higher than the 80 percentile where kept, other values where set to zero. Using this profile we ran a grid

search optimisation (See previous paragraph) to obtain MRT_0 and RFD_0 that fitted best experimental one. Then a neural network N_1 was trained to predict the initiation profile $I_0 + r$ (r being the amount of random activation obtained from the grid search optimisation part) from the simulated MRT_0 and RFD_0 (See next paragraph for details on the neural network parameters). N_1 was then applied on experimental MRT and RFD predicting the I_1 profile. Then I_1 was used as input of a simulation to obtain MRT_1 and RFD_1. We then computed the correlation with experimental MRT and RFD. To summarise in this iteration we used MRT_0 and RFD_0 as input to produce I_1, MRT_1 and RFD_1. We then reiterated the process twice using successively MRT_1 and RFD_1 as input and then the obtained MRT_2 and RFD_2 as input to produce MRT_3 and RFD_3. we reitirated once more the procedure and stopped as it did not improved the correlations. All the code to reproduce the steps here are detailed at (https://github.com/organic-chemistry/repli1D)

Neural network training

The inputs of the networks is a window of size 2005 kb (401 bins of 5 kb) with both MRT and RFD, an the ouptut is the initiation signal at the center of the window. The RFD is smoothed over 50 kb by a rolling window. We used a three layer convolutional neural network with kernel length 10 and filter size 15 and a relu activation. Each layers is followed by a dropout layer at a value of 1 %. The last convolutional layer is followed by a maxpooling of kernel size 2. Then the resulting vector go through a dense layer with sigmoid activation and output size 1, meaning that the 2005 kb window allow to compute the probability of activation at its center. We use a binary cross entropy loss and the layer is trained with the adadelta algorithm. The following code has been implemented using keras.

We used chromosome 3 to 23 for the training, chromosome 2 for validation and chromosome 1 for testing.

To make the network more robust to experimental noise, we randomly add noise to the input RFD profile: 1% of the bin are assigned a random value between -1 and 1.

Analytical extraction of n_e from ΔRFD and MRT data

Eq (7) requires that ΔRFD be estimated. Before computing ΔRFD in human, we smoothed RFD using a running average window of 15 kb (3 bins). Then we computed ΔRFD between consecutive 5kb windows. The 15% top values are selected and the other bins are set to 0. Then the non zero bin are divided by $1/MRT$ or by e^{-6MRT} .

Selecting RFD peaks

We ran a peak detection algorithm over ΔRFD at a 5 kb resolution using scipy routine find_peaks with parameter width=3 and height=0.035. Like this we selected peak larger or equal to 3*5kb and higher than 0.035 delta RFD/5kb. One can visualise the selected peaks as well as ΔRFD and I_3 for comparison in Fig S17. This yielded 9644 peaks for GM06990 , 14590 peaks for Hela and 9787 peaks for K562.

References

[1] Jean-Michel Arbona et al. "The eukaryotic bell-shaped temporal rate of DNA replication origin firing emanates from a balance between origin activation and passivation". In: *eLife* 7 (June 2018). DOI: 10.7554/elife.35192.

[2] Benjamin Audit et al. "Multiscale analysis of genome-wide replication timing profiles using a wavelet-based signal-processing algorithm". In: *Nature Protocols* 8.1 (Dec. 2012), pp. 98–110. DOI: 10.1038/nprot.2012.145.

[3] A Baker and J Bechhoefer. "Inferring the spatiotemporal DNA replication program from noisy data". In: *Physical Review E* 89.3 (2014), p. 032703.

[4] Antoine Baker et al. "Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines". eng. In: *PLoS Comput. Biol.* 8.4 (2012), e1002443. DOI: 10.1371/journal.pcbi.1002443. URL: http://dx.doi.org/10.1371/journal.pcbi.1002443.

[5] Alina Bazarova et al. "Bayesian inference of origin firing time distributions, origin interference and licencing probabilities from Next Generation Sequencing data". In: *Nucleic acids research* 47.5 (2019), pp. 2229–2243.

[6] J. Bechhoefer and N. Rhind. "Replication timing and its emergence from stochastic processes". In: *Trends Genet.* 28.8 (2012), pp. 374–381. DOI: 10.1016/j.tig.2012.03.011.

[7] Emilie Besnard et al. "Unraveling cell type–specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs". In: *Nature Structural & Molecular Biology* 19.8 (July 2012), pp. 837–844. DOI: 10.1038/nsmb.2339.

[8] Marion Blin et al. "DNA molecular combing-based replication fork directionality profiling". In: *Nucleic Acids Research* (2021).

[9] Chun-Long Chen et al. "Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes". In: *Genome research* 20.4 (2010), pp. 447–457.

[10] Clara Collart et al. "Titration of four replication factors is essential for the Xenopus laevis midblastula transition". In: *Science* 341.6148 (2013), pp. 893–896.

[11] Shankar P. Das et al. "Replication timing is regulated by the number of MCMs loaded at origins." In: *Genome Res.* 25 (12 Dec. 2015), pp. 1886–1892. ISSN: 1549-5469. DOI: 10.1101/gr.195305.115.

[12] Melvin DePamphilis and Stephen Bell. *Genome Duplication: concepts, mechanisms, evolution and disease.* Garland Science, New York, 2010.

[13] Max E Douglas and John FX Diffley. "Replication timing: the early bird catches the worm". In: *Current Biology* 22.3 (2012), R81–R82.

[14] Max E Douglas et al. "The mechanism of eukaryotic CMG helicase activation". In: *Nature* 555.7695 (2018), pp. 265–268.

[15] Melissa C Edwards et al. "MCM2–7 Complexes Bind Chromatin in a Distributed Pattern Surrounding the Origin Recognition Complex inXenopus Egg Extracts". In: *Journal of Biological Chemistry* 277.36 (2002), pp. 33049–33057.

[16] Cecile Evrin et al. "A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication". In: *Proceedings of the National Academy of Sciences* 106.48 (2009), pp. 20240–20245.

[17] Eric J Foss et al. "Chromosomal Mcm2-7 distribution and the genome replication program in species from yeast to humans". In: *PLoS Genetics* 17.9 (2021), e1009714.

[18] Daniel T. Gillespie. "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions". In: *J. Comput. Phys.* 22.4 (1976), pp. 403–434. DOI: 10.1016/0021-9991(76)90041-3. URL: http://www.sciencedirect.com/science/article/pii/0021999176900413.

[19] Y. Gindin et al. "A chromatin structure-based model accurately predicts DNA replication timing in human cells". In: *Molecular Systems Biology* 10.3 (Mar. 2014), pp. 722–722. DOI: 10.1002/msb.134859.

[20] Arach Goldar, Marie-Claude Marsolier-Kergoat, and Olivier Hyrien. "Universal temporal profile of replication origin activation in eukaryotes". eng. In: *PLoS One* 4.6 (2009), e5899. DOI: 10.1371/journal.pone.0005899. URL: http://dx.doi.org/10.1371/journal.pone.0005899.

[21] Arach Goldar et al. "A dynamic stochastic model for DNA replication initiation in early embryos". eng. In: *PLoS One* 3.8 (2008), e2919. DOI: 10.1371/journal.pone.0002919. URL: http://dx.doi.org/10.1371/journal.pone.0002919.

[22] Guillaume Guilbaud et al. "Evidence for Sequential and Increasing Activation of Replication Origins along Replication Timing Gradients in the Human Genome". In: *PLoS Computational Biology* 7.12 (Dec. 2011). Ed. by Christopher E. Pearson, e1002322. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1002322.

[23] Angela T Hahn, Joshua T Jones, and Tobias Meyer. "Quantitative analysis of cell cycle phase durations and PC12 differentiation using fluorescent biosensors". In: *Cell cycle* 8.7 (2009), pp. 1044–1052.

[24] Joyce L Hamlin, Larry D Mesner, and Pieter A Dijkwel. "A winding road to origin discovery". In: *Chromosome research* 18.1 (2010), pp. 45–61.

[25] R Scott Hansen et al. "Sequencing newly replicated DNA reveals widespread plasticity in human replication timing". In: *Proceedings of the National Academy of Sciences* 107.1 (2010), pp. 139–144.

[26] Kevin J Harvey and John Newport. "CpG methylation of DNA restricts prereplication complex assembly in Xenopus egg extracts". In: *Molecular and cellular biology* 23.19 (2003), pp. 6769–6779.

[27] Magali Hennion et al. "FORK-seq: replication landscape of the Saccharomyces cerevisiae genome by nanopore sequencing". In: *Genome biology* 21.1 (2020), pp. 1–25.

[28] Magali Hennion et al. "Mapping DNA replication with nanopore sequencing". In: *bioRxiv* (2018). DOI: 10.1101/426858. eprint: https://www.biorxiv.org/content/early/2018/09/26/426858.full.pdf. URL: https://www.biorxiv.org/content/early/2018/09/26/426858.

[29] Michelle L Hulke, Dashiell J Massey, and Amnon Koren. "Genomic methods for measuring DNA replication dynamics". In: *Chromosome Research* 28.1 (2020), pp. 49–67.

[30] Olivier Hyrien. "Peaks cloaked in the mist: the landscape of mammalian replication origins". In: *Journal of Cell Biology* 208.2 (2015), pp. 147–160.

[31] Nina Kirstein et al. "Human ORC/MCM density is low in active genes and correlates with replication time but does not delimit initiation zones". In: *Elife* 10 (2021), e62161.

[32] Amnon Koren et al. "Genetic variation in human DNA replication timing". In: *Cell* 159.5 (2014), pp. 1015–1026.

[33] Anne Letessier et al. "Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site". In: *Nature* 470.7332 (2011), pp. 120–123.

[34] D. Löb et al. "3D replicon distributions arise from stochastic initiation and domino-like DNA replication progression". In: *Nat. Commun.* 7 (Apr. 2016), p. 11207. ISSN: 2041-1723. DOI: 10.1038/ncomms11207.

[35] Haizhen Long et al. "H2A. Z facilitates licensing and activation of early replication origins". In: *Nature* 577.7791 (2020), pp. 576–581.

[36] Davide Mantiero et al. "Limiting replication initiation factors execute the temporal programme of origin firing in budding yeast". In: *The EMBO journal* 30.23 (2011), pp. 4805–4814.

[37] Larry D Mesner et al. "Bubble-seq analysis of the human genome reveals distinct chromatin-mediated mechanisms for regulating early-and late-firing origins". In: *Genome research* 23.11 (2013), pp. 1774–1788.

[38] Thomas CR Miller et al. "Mechanism of head-to-head MCM double-hexamer formation revealed by cryo-EM". In: *Nature* 575.7784 (2019), pp. 704–710.

[39] Benoit Miotto, Zhe Ji, and Kevin Struhl. "Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers". en. In: *Proceedings of the National Academy of Sciences* 113.33 (Aug. 2016), E4810–E4819. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.1609060113. URL: http://www.pnas.org/content/113/33/E4810 (visited on 03/23/2018).

[40] Alessandro PS de Moura et al. "Mathematical modelling of whole chromosome replication". In: *Nucleic Acids Res.* 38.17 (2010), pp. 5623–5633.

[41] Carolin A Müller et al. "Capturing the dynamics of genome replication on individual ultra-long nanopore sequence reads". In: *Nature methods* 16.5 (2019), pp. 429–436.

[42] Carolin A Müller et al. "The dynamics of genome replication using deep sequencing". In: *Nucleic Acids Research* 42.1 (2014), e3–e3.

[43] Pedro D Pereira et al. "Quantification of cell cycle kinetics by EdU (5-ethynyl-2′deoxyuridine)-coupled-fluorescence-intensity analysis". In: *Oncotarget* 8.25 (2017), p. 40514.

[44] Nataliya Petryk et al. "MCM2 promotes symmetric inheritance of modified histones during DNA replication". In: *Science* 361.6409 (2018), pp. 1389–1392.

[45] Nataliya Petryk et al. "Replication landscape of the human genome". en. In: *Nat Commun* 7 (Jan. 2016), p. 10208. ISSN: 2041-1723. DOI: 10.1038/ncomms10208. URL: http://www.nature.com/ncomms/2016/160111/ncomms10208/full/ncomms10208.html (visited on 01/04/2017).

[46] Franck Picard et al. "The Spatiotemporal Program of DNA Replication Is Associated with Specific Combinations of Chromatin Marks in Human Cells". In: *PLoS Genetics* 10.5 (May 2014). Ed. by Bas van Steensel, e1004282. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1004282.

[47] Dirk Remus et al. "Concerted loading of Mcm2–7 double hexamers around DNA during DNA replication origin licensing". In: *Cell* 139.4 (2009), pp. 719–730.

[48] Jan Marten Schmidt and Franziska Bleichert. "Structural mechanism for replication origin binding and remodeling by a metazoan origin recognition complex and its co-loader Cdc6". In: *Nature communications* 11.1 (2020), pp. 1–17.

[49] Nozomi Sugimoto et al. "Genome-wide analysis of the spatiotemporal regulation of firing and dormant replication origins in human cells". In: *Nucleic acids research* 46.13 (2018), pp. 6683–6696.

[50] Seiji Tanaka et al. "Origin association of Sld3, Sld7, and Cdc45 proteins is a key step for determination of origin-firing timing". In: *Current Biology* 21.24 (2011), pp. 2055–2063.

[51] Hervé Técher et al. "Replication dynamics: biases and robustness of DNA fiber analysis". In: *Journal of molecular biology* 425.23 (2013), pp. 4845–4855.

[52] Robert E Thurman et al. "Identification of higher-order functional domains in the human ENCODE regions". In: *Genome research* 17.6 (2007), pp. 917–927.

[53] Anthony Tubbs et al. "Dual roles of poly (dA: dT) tracts in replication initiation and fork collapse". In: *Cell* 174.5 (2018), pp. 1127–1142.

[54] Weitao Wang et al. "Genome-wide mapping of human DNA replication by optical replication mapping supports a stochastic model of eukaryotic replication". In: *Molecular Cell* (2021).

[55] Tom Serge Weber et al. "Quantifying the length and variance of the eukaryotic cell cycle phases by a stochastic model and dual nucleoside pulse labelling". In: *PLoS computational biology* 10.7 (2014), e1003616.

[56] Michael Christian Weis. *Computational models of the mammalian cell cycle*. Case Western Reserve University, 2012.

[57] RJ Widrow et al. "Very late DNA replication in the human cell cycle". In: *Proceedings of the National Academy of Sciences* 95.19 (1998), pp. 11246–11250.

[58] Philip G Wong et al. "Cdc45 limits replicon usage from a low density of preRCs in mammalian cells". In: *PloS one* 6.3 (2011), e17533.

[59] Xia Wu et al. "Developmental and cancer-associated plasticity of DNA replication preferentially targets GC-poor, lowly expressed and late-replicating regions". In: *Nucleic Acids Research* 46.19 (Sept. 2018), pp. 10157–10172. ISSN: 0305-1048. DOI: 10.1093/nar/gky797. eprint: https://academic.oup.com/nar/article-pdf/46/19/10157/26333861/gky797_supplemental_files.pdf. URL: https://doi.org/10.1093/nar/gky797.

[60] Scott Cheng-Hsin Yang, Nicholas Rhind, and John Bechhoefer. "Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing". In: *Mol. Syst. Biol.* 6.1 (2010), p. 404.

[61] Peiyao A Zhao, Takayo Sasaki, and David M Gilbert. "High-resolution Repli-Seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells". In: *Genome biology* 21.1 (2020), pp. 1–20.

Supporting information

PODLS	K562 MRT	K562 RFD	GM MRT	GM RFD	Hela MRT	Hela RFD	*S. cer.* MRT	*S. cer.* RFD
I_4	0.98	0.91	0.98	0.92	0.98	0.85	0.96	0.91
I_3	0.98	0.91	0.98	0.92	0.98	0.85	0.94	0.90
I_2	0.98	0.91	0.98	0.90	0.97	0.84	0.87	0.84
I_1	0.95	0.90	0.95	0.87	0.93	0.82	0.80	0.83
I_0	0.79	0.80	0.69	0.76	0.77	0.74	0.60	0.71

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).

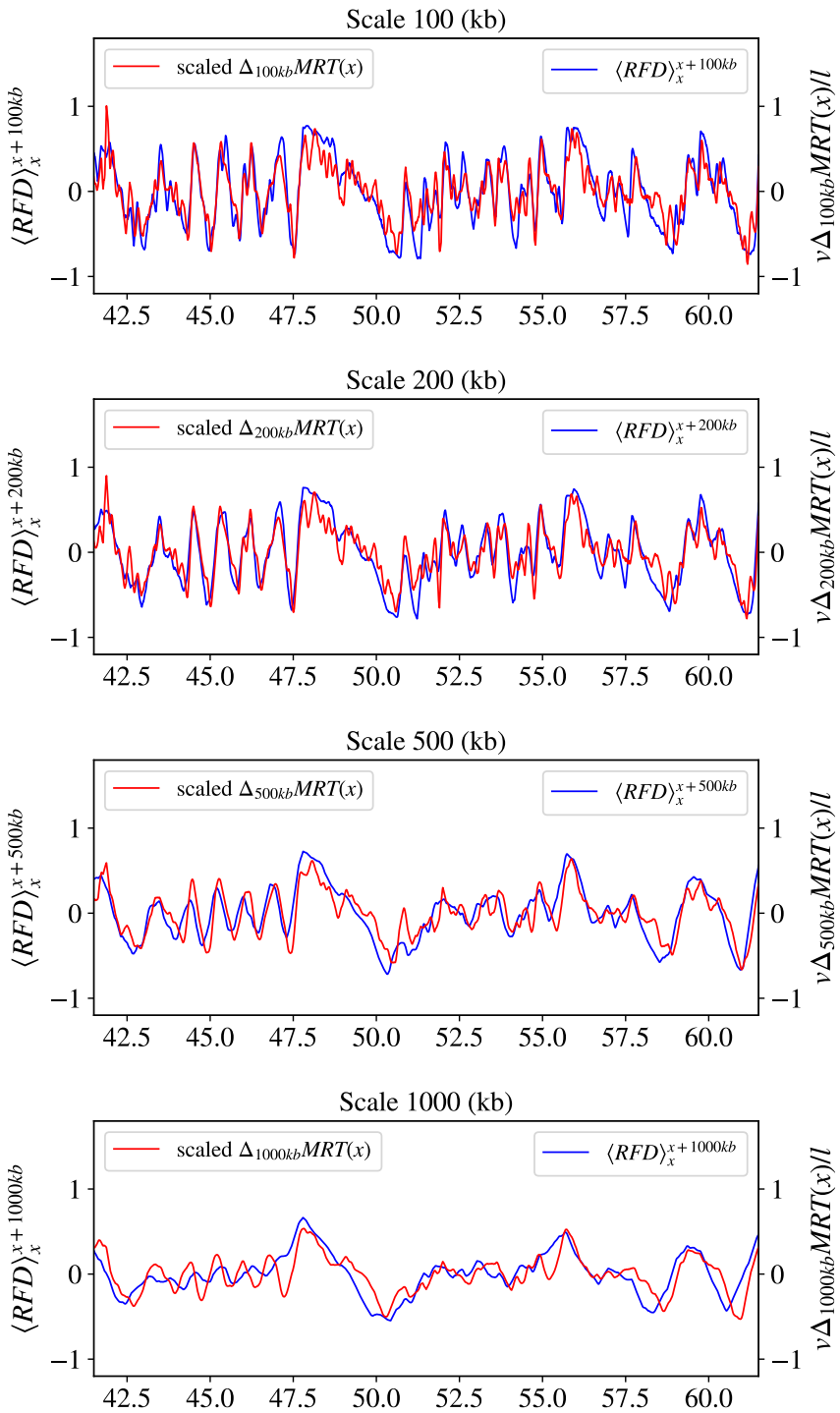

Fig S1. Comparison between $\langle RFD\rangle_x^{x+l}$ (blue) and $\frac{v}{l}\Delta_l MRT(x)$ (red) (Eq. (2)) assuming $T_S = 12$ at different scale l. From (top) to (bottom), $l = 100$ kb, 200 kb, 500 kb and 1000 kb and fork speed values v are taken from Fig. 2B. The same 20 Mb region of chromosme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

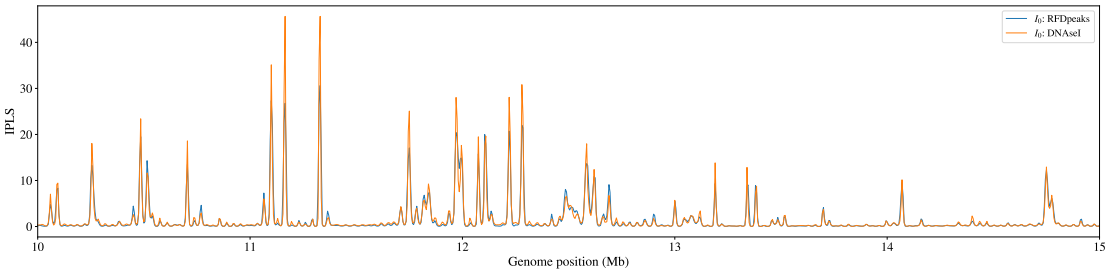

Fig S2. Comparison of K562 potential origin density landscape I_4 optimized starting from the I_0 PODLS derived from DNase I HS peaks (orange) or experimental RFD peaks (blue). The obtained I_4 are almost identical (PCC = 0.94).

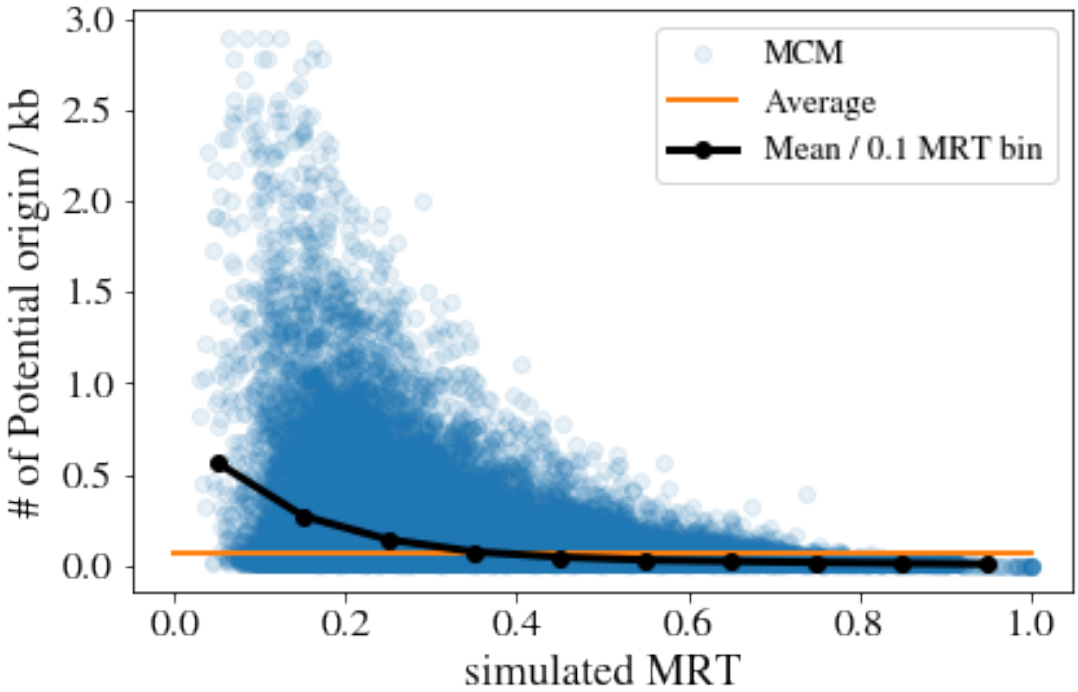

Fig S3. Number of potential origin / kb for K562 (computed for each bin)

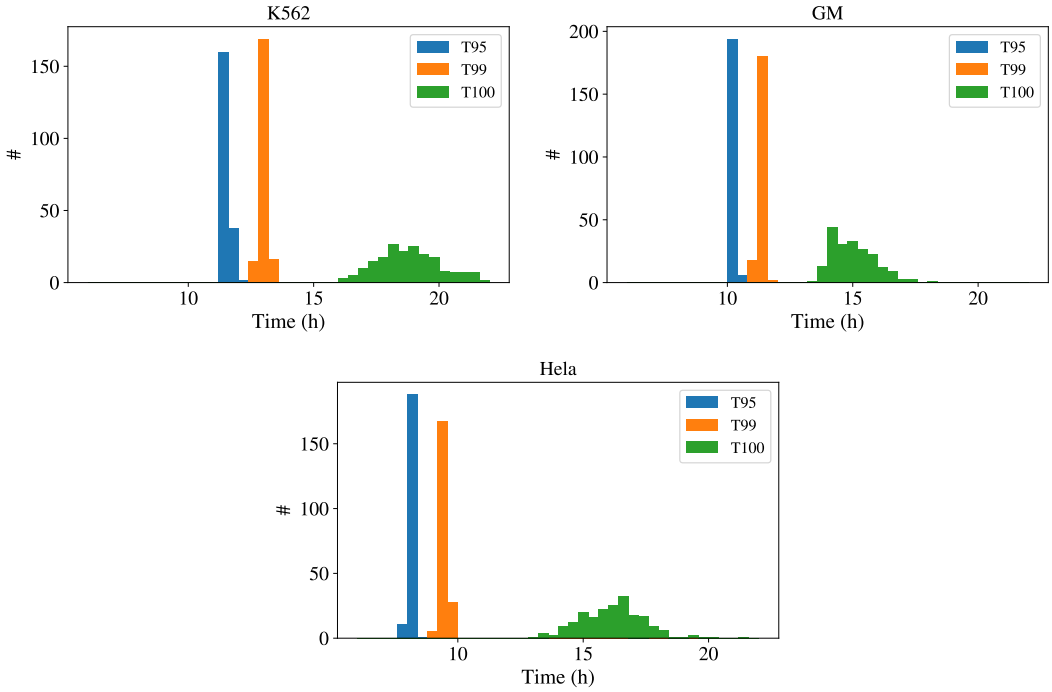

Fig S4. Distribution of replication time

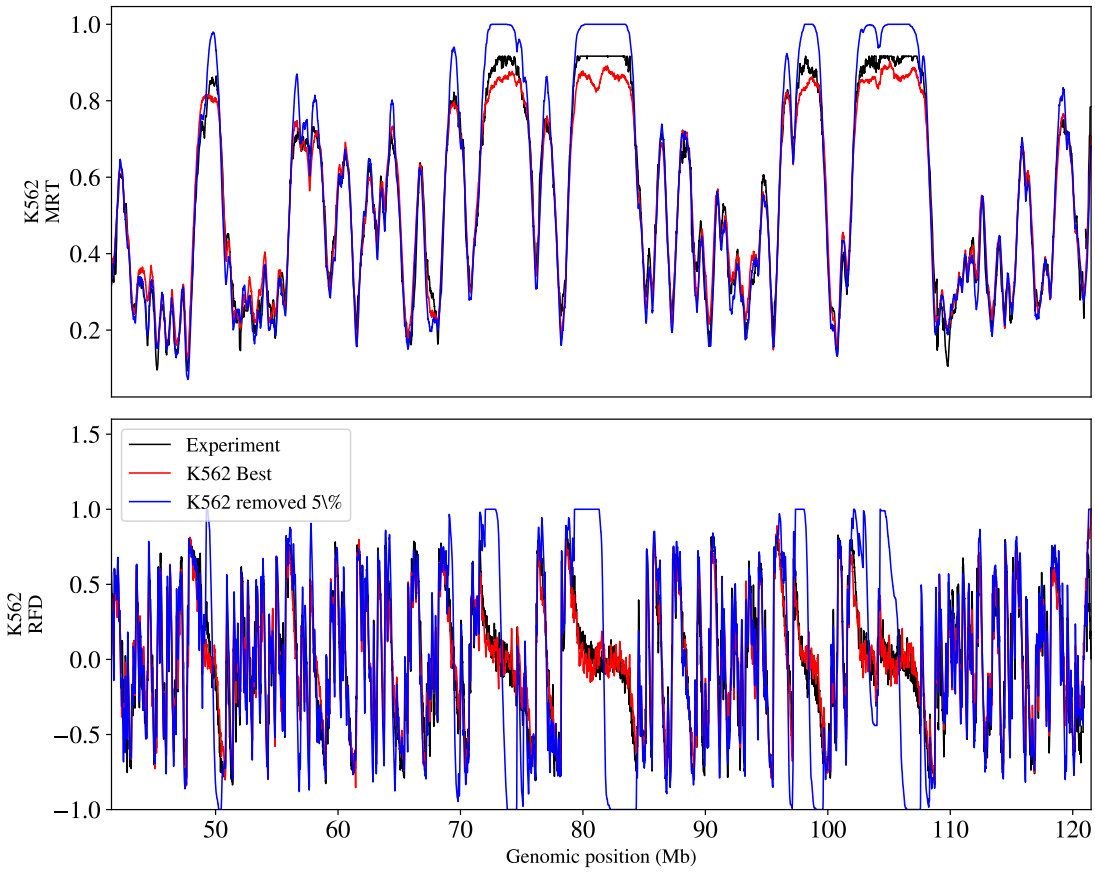

Fig S5. (Top) Comparison of experimental MRT with simulated MRT using the optimized PODLS I_4 (red) and after setting to zero the bins with the lowest I_4 values corresponding to 5% of the total activations (blue). (Bottom) Same comparison for the RFD profiles.

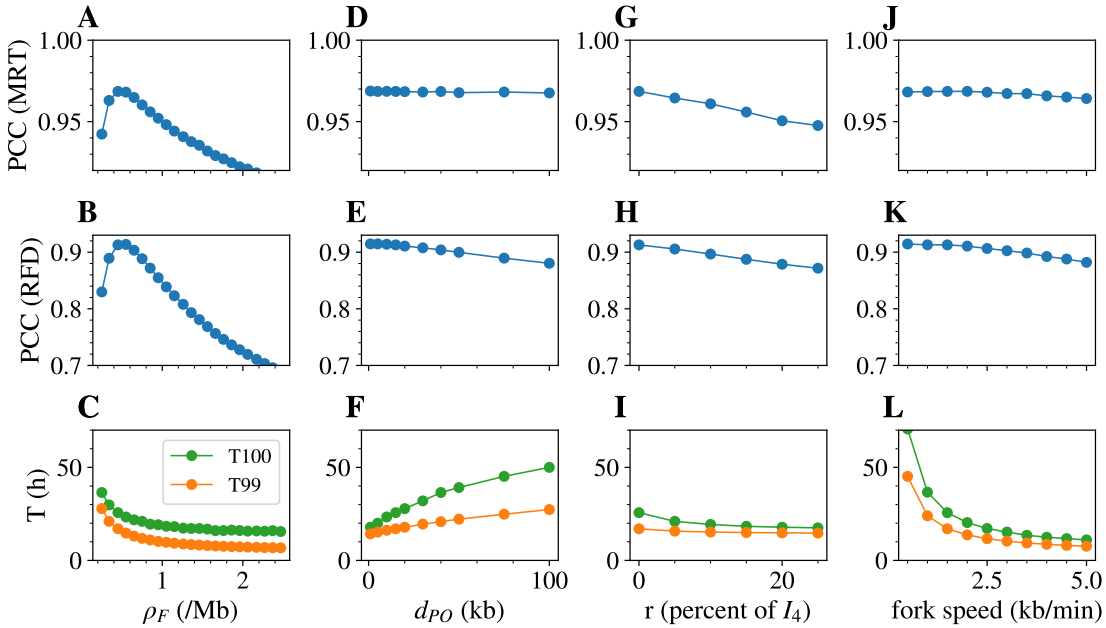

Fig S6. Effect of single parameter variation on measurable targets in GM. Effect of the density of firing factors ρ_F (A,B,C); the average distance between potential origins d_{PO} (D,E,F) ; the percent of random initiation r (G,H,I); the fork speed v (J,K,L), on the Pearson Correlation Coefficient (PCC)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

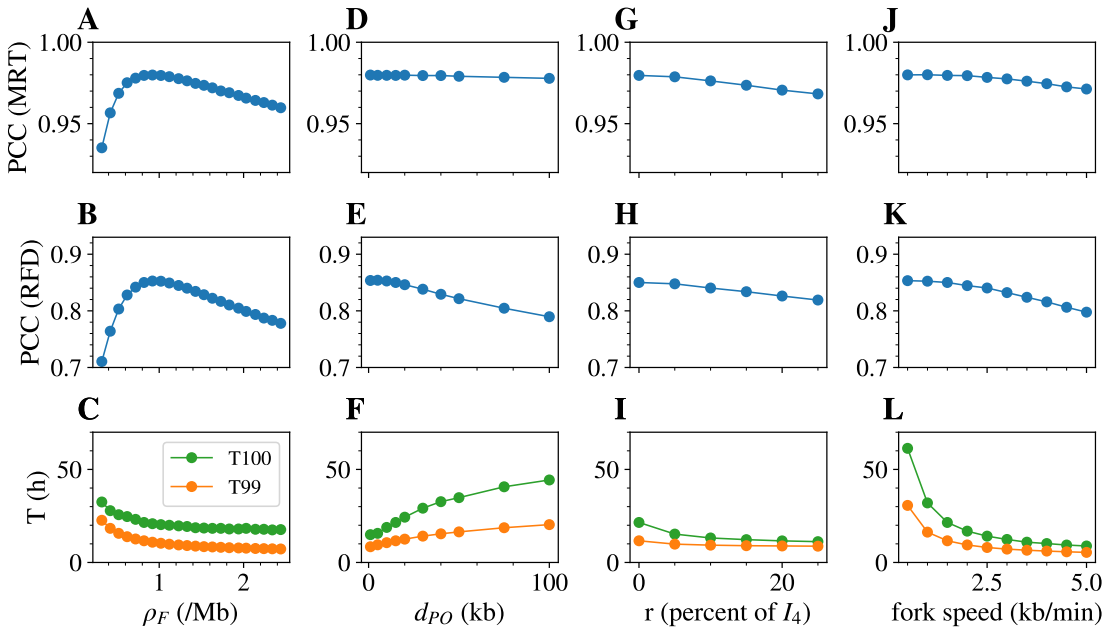

Fig S7. Effect of single parameter variation on measurable targets in Hela. Effect of the density of firing factors ρ_F (A,B,C); the average distance between potential origins d_{PO} (D,E,F) ; the percent of random initiation r (G,H,I); the fork speed v (J,K,L), on the Pearson Correlation Coefficient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

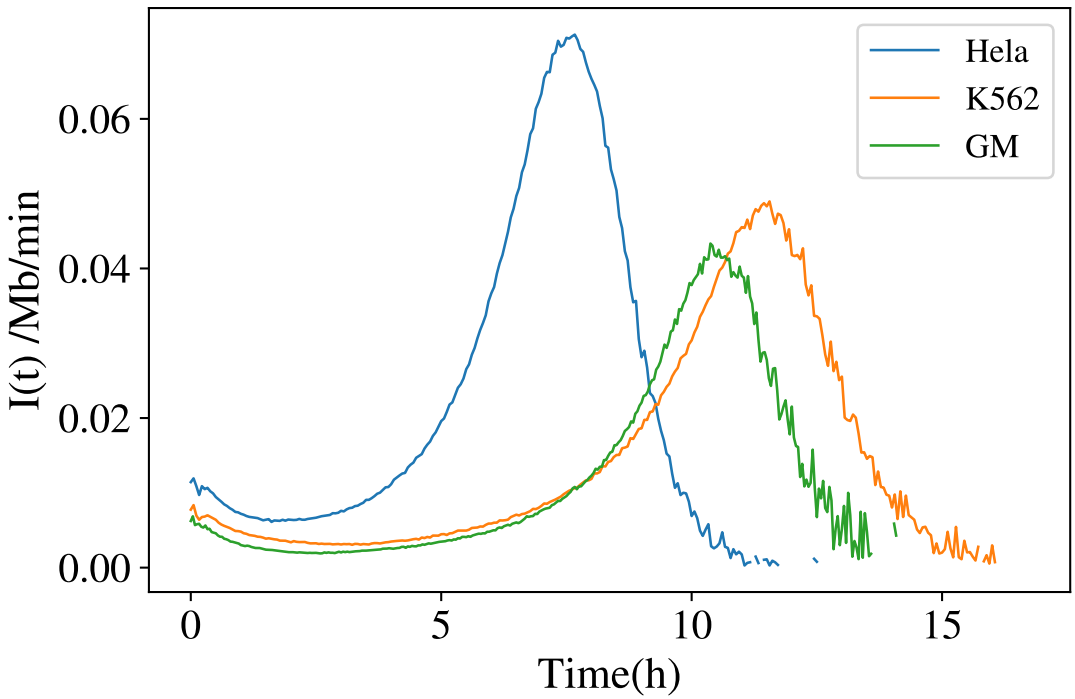

Fig S8. Probability of initiation per length of unreplicated DNA for the three cell lines

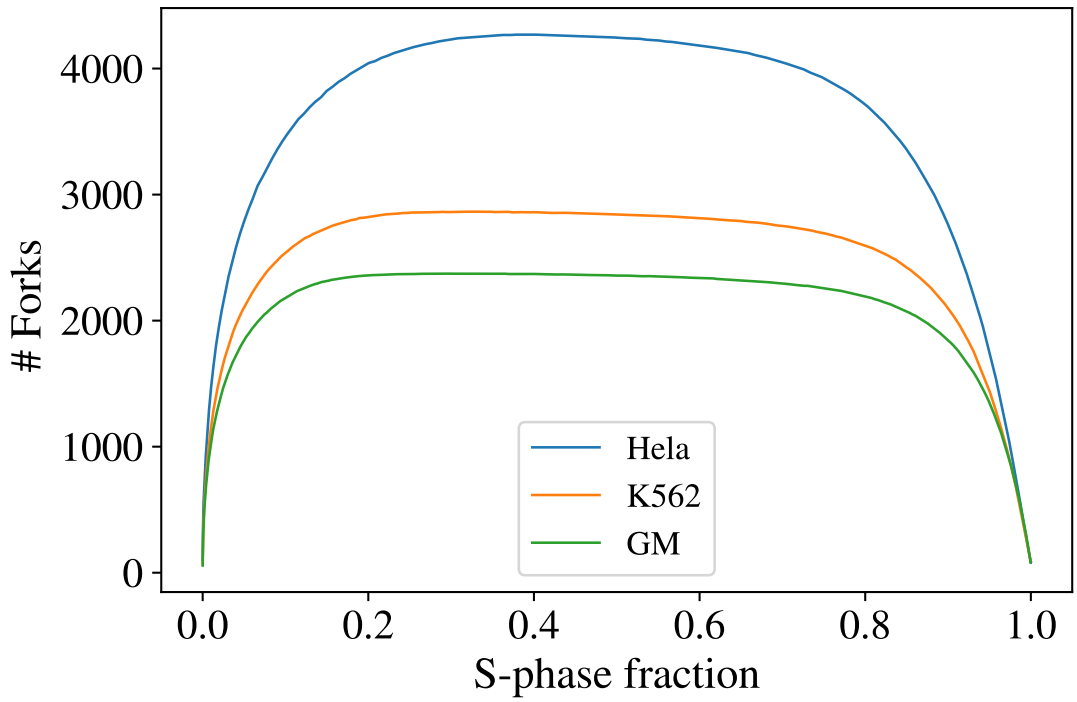

Fig S9. Active fork number as a function of S-phase fraction for the three cell lines

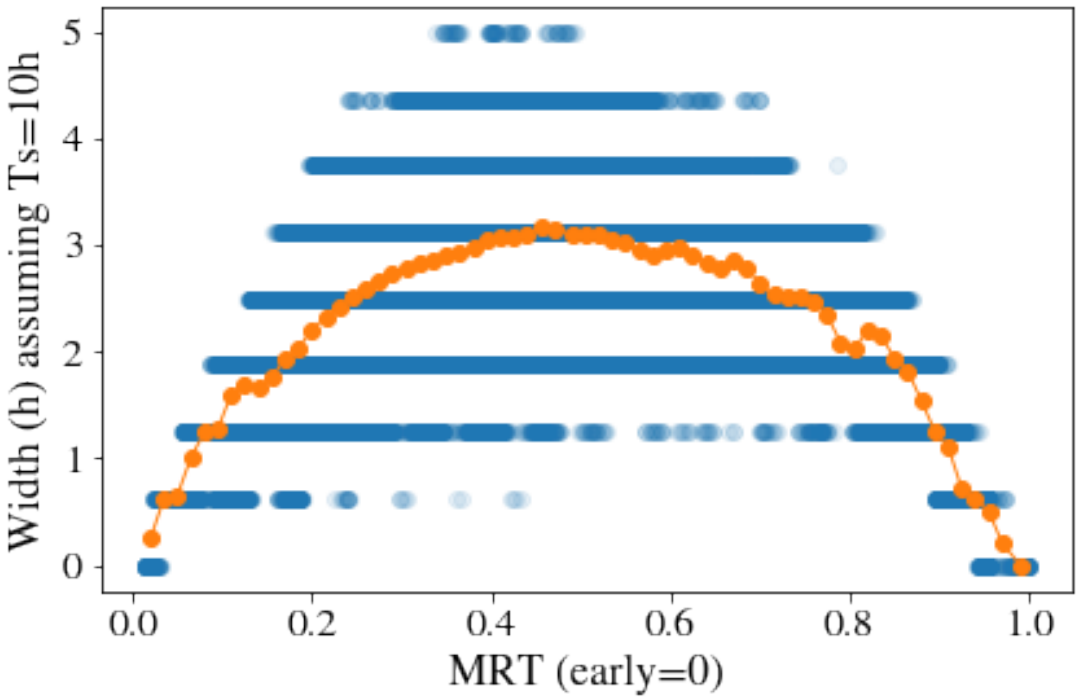

Fig S10. Assuming a T=10h S phase as in (Gilbert)

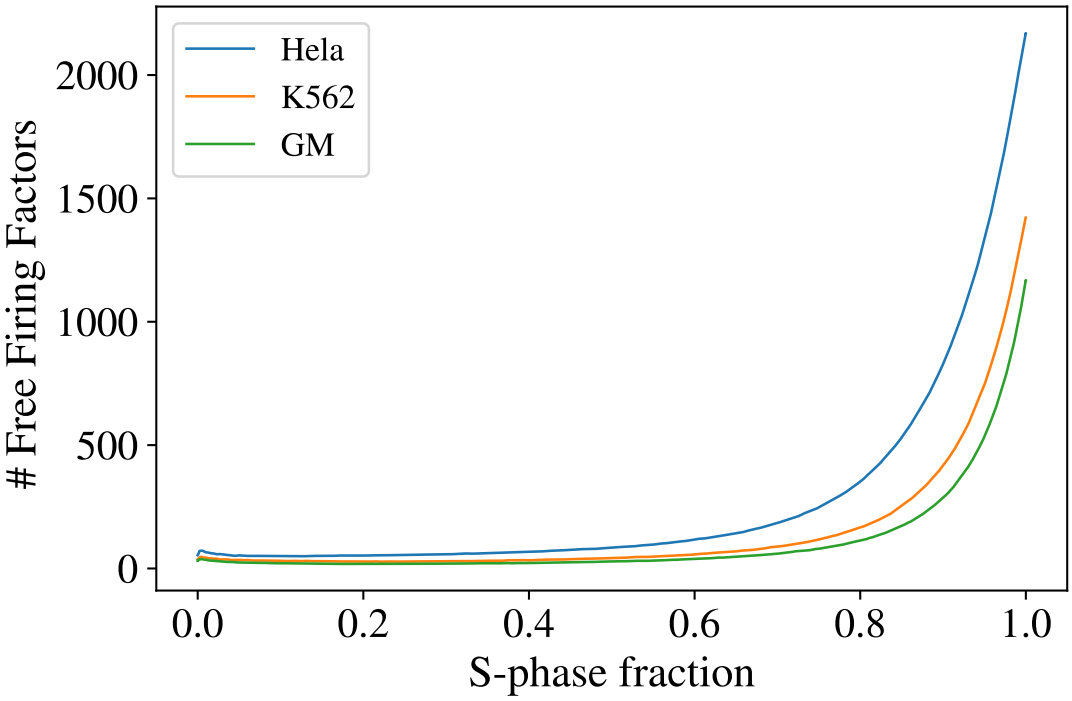

Fig S11. Number of Free firing factors as a function of the S-phase fraction for the three cell lines.

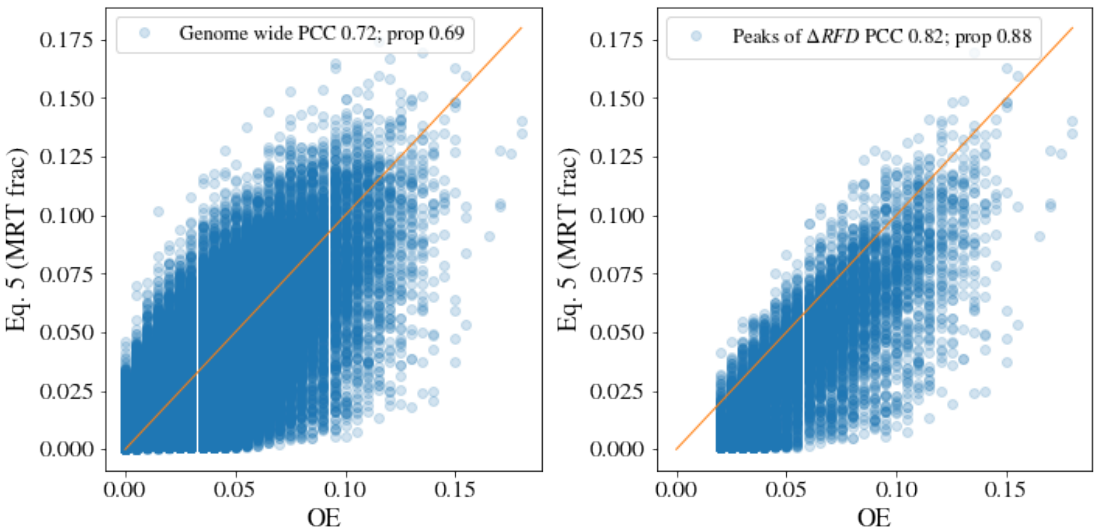

Fig S12. OE vs eq (5) for the MRT computed in fraction of S phase, genome wide (left) or restricted to the Peaks of ΔRFD (right)

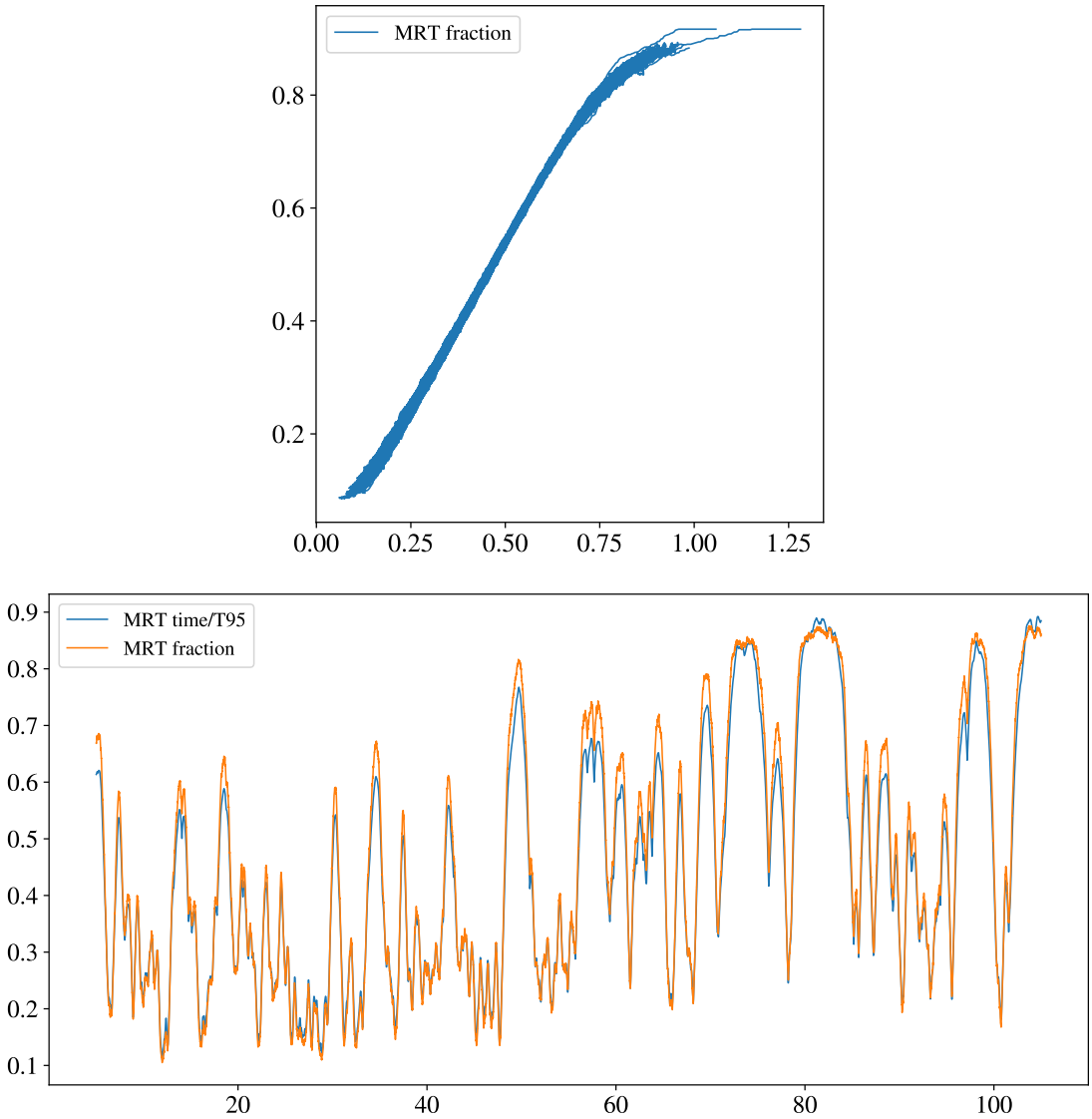

Fig S13. Comparison of MRT computed as a fraction along the S-phase or with real time (top) genome wide. (Bottom) for part of chromosome 1

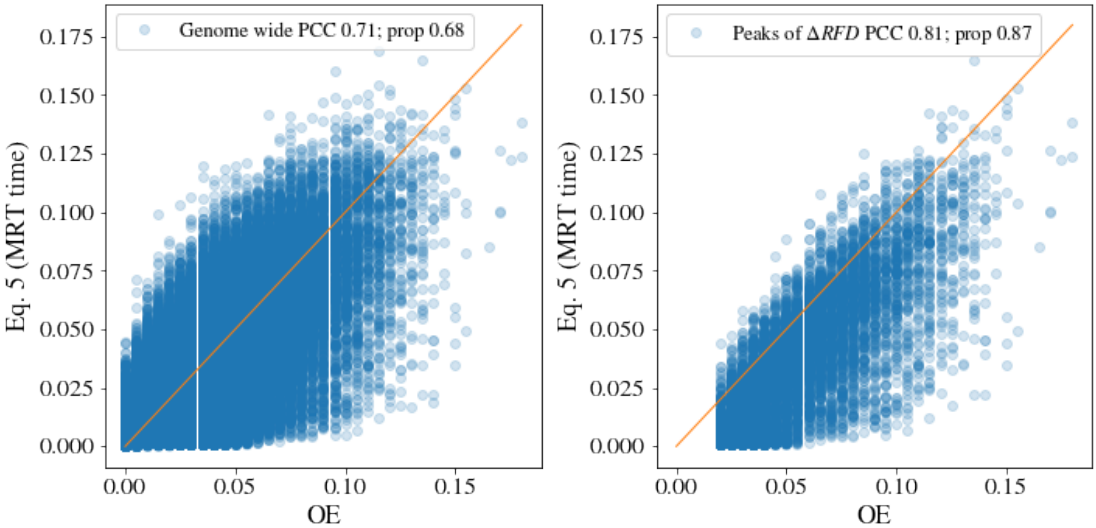

Fig S14. OE vs eq (5) for real MRT genome wide (left) or restricted to the peaks of ΔRFD (right)

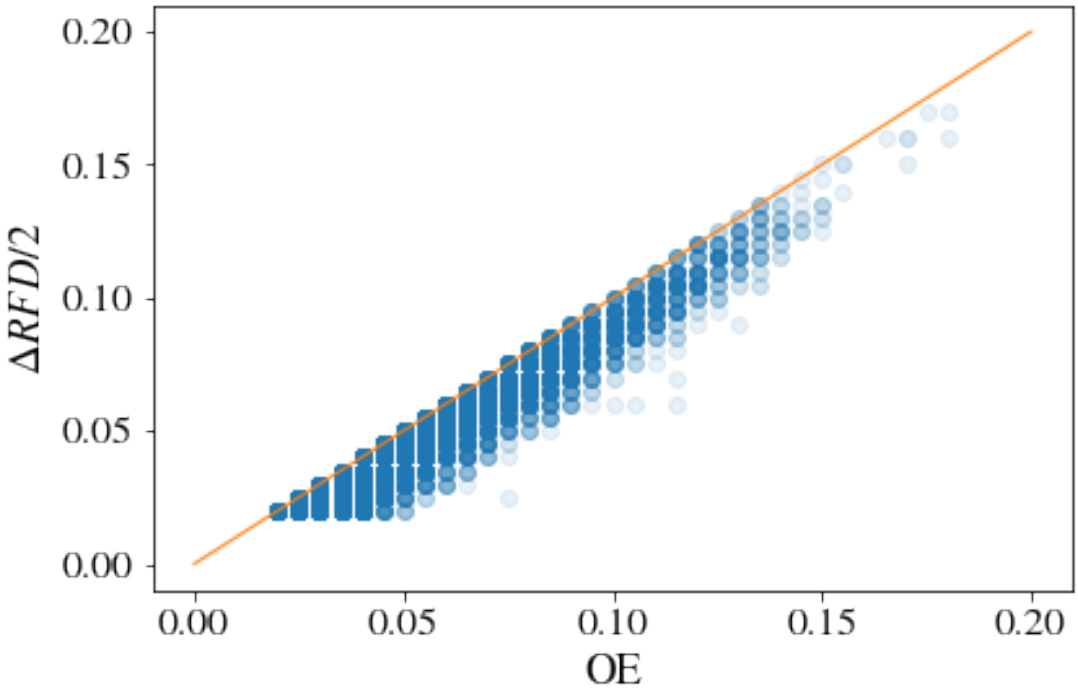

Fig S15. OE vs $\Delta RFD/2$ restricted to peaks of ΔRFD

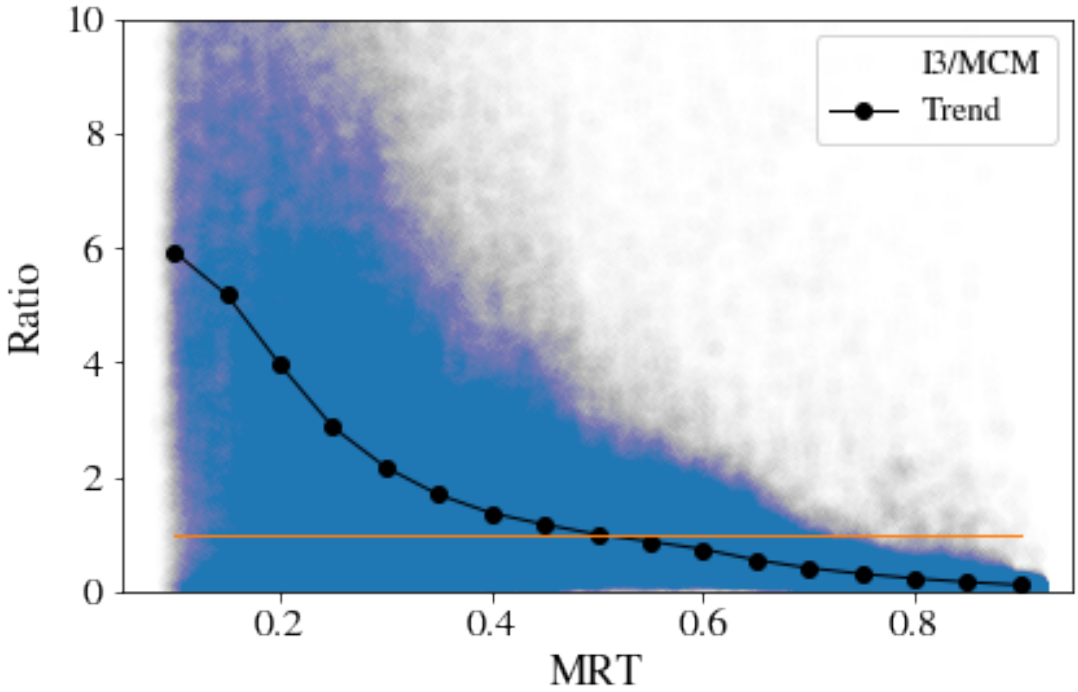

Fig S16. Ratio of I_4/MCM . Both signals where first smoothed at 50 kb as MCM experimental signal [17] is noisy. Then the ratio was normalized so that the median value over all the genome is one. The trend indicate the median value of the ratio by 0.05 MRT steps.

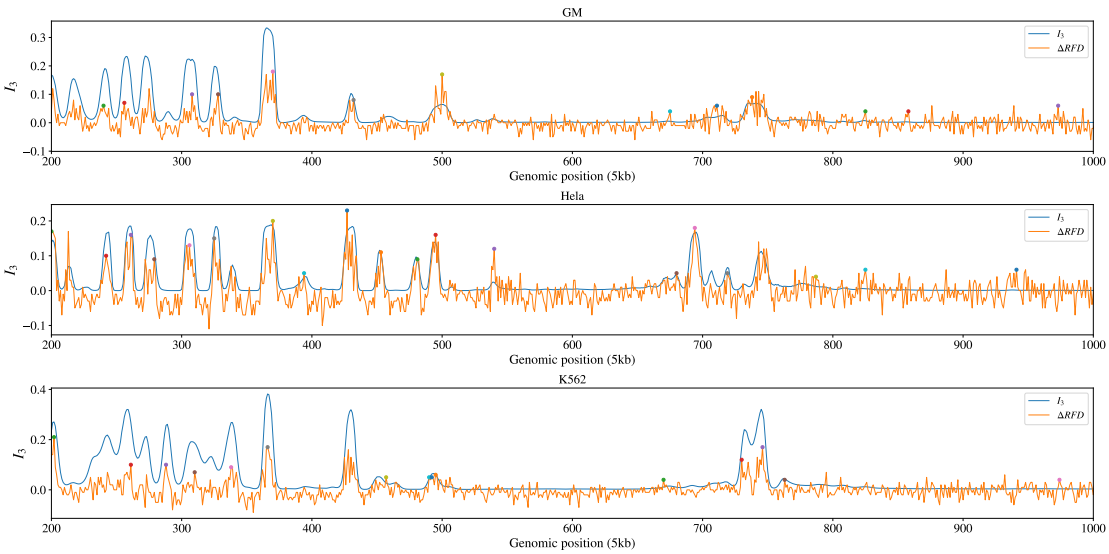

Fig S17. Peak detection of ΔRFD for the three different cell lines overlayed with I_3