bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DNA replication timing reveals genome-wide features of transcription and fragility

Francisco Berkemeier [1,2,✉], **Peter R. Cook** [3], **and Michael A. Boemo** [1,2,✉]

[1]Department of Pathology, University of Cambridge, CB2 1QP, Cambridge, United Kingdom
[2]Department of Genetics, University of Cambridge, CB2 3EH, Cambridge, United Kingdom
[3]Sir William Dunn School of Pathology, University of Oxford, OX1 3RE, Oxford, United Kingdom

✉Corresponding authors: Francisco Berkemeier (fp409@cam.ac.uk), and Michael A. Boemo (mb915@cam.ac.uk)

DNA replication in humans requires precise regulation to ensure accurate genome duplication and maintain genome integrity. A key indicator of this regulation is replication timing, which reflects the interplay between origin firing and fork dynamics. We present a high-resolution (1-kilobase) mathematical model that maps firing rate distributions to replication timing profiles across various cell lines, validated using Repli-seq data. The model effectively captures genome-wide replication patterns while identifying local discrepancies. Notably, regions where the model and data diverge often overlap with fragile sites and large genes, highlighting the influence of genomic architecture on replication dynamics. Conversely, regions of high concordance are associated with open chromatin and active promoters, where elevated firing rates facilitate timely fork progression and reduce replication stress. By establishing these correlations, our model provides a valuable framework for exploring the structural interplay between replication timing, transcription, and chromatin organisation, offering new insights into the mechanisms underlying replication stress and its implications for genome stability and disease.

DNA Replication | Origin Firing | Fragile Sites | Transcription | Mathematical Model

Introduction

Accurate DNA replication is essential for faithfully duplicating genetic information, ensuring its preservation for future generations (Gefter, 1975). In humans, replication occurs during the S phase when multiple discrete chromosomal sites, termed origins of replication (Leonard and Méchali, 2013), "fire" to initiate bidirectional replication forks—molecular machines that traverse the chromosome and replicate DNA (Waga and Stillman, 1998). These forks move in opposite directions, progressing until they encounter another fork, reach the chromosome end (Figure 1a), or face an obstacle (e.g., a bound protein or transcription complex; Mirkin and Mirkin (2007)). Intriguingly, each origin fires stochastically so firing sites and times differ from cell to cell. Despite this apparent randomness, consistent trends emerge so that different cell types have characteristic firing profiles (Rhind and Gilbert, 2013).

Replication timing is a critical marker of replication fidelity, reflecting the time it takes for a specific locus either to fire (if an origin) or to be passively replicated by an incoming fork. These timing profiles are closely associated with various chromatin structures (Marchal et al., 2019), as well as gene expression (Müller and Nieduszynski, 2017; Maric and Prioleau, 2010; Kogoma, 1997) and replication stresses (Briu et al., 2021). Furthermore, replication timing is linked to genetic variation (Koren et al., 2014) and cancer (where late or delayed replication often correlates with increased genomic instability; Woo and Li (2012); Donley and Thayer (2013)). Of particular interest are fragile sites, regions that are especially vulnerable to breakage due to replication stress, and are often found in late-replicating regions (Laird et al., 1987; Sinai and Kerem, 2018). These sites, and the long genes found within them, are often hotspots for the chromosomal rearrangements and deletions that arise in cancers and other genetic diseases (Smith et al., 2006).

Replication, transcription, and chromatin organisation are also intricately inter-connected, with each influencing the other (Ehrenhofer-Murray, 2004; Sequeira-Mendes et al., 2009; Turner and Woodworth, 2001; Tabancay Jr and Forsburg, 2006; Kadonaga, 1998). In particular, chromatin remodelling regulates the accessibility of regulatory factors, influencing both gene expression and the replication process. Open chromatin is strongly linked to transcriptional activity and plays a crucial role in replication timing (Guilbaud et al., 2011; Audit et al., 2009).

Although associations between genomic features are well-established, identifying site-specific or context-dependent differences remains a challenge. Experimental approaches often struggle to isolate individual variables, limiting our ability to disentangle the interplay between replication and other processes. To address these gaps, we develop a stochastic model that maps origin firing rates to replication timing, capturing variability across cell populations. By integrating data from RNA-seq (Marguerat and Bähler, 2010), ChIP-seq (Pepke et al., 2009), GRO-seq (Lopes et al., 2017), and a database of fragile sites (HumCFS; Kumar et al. (2019)), we provide a framework to explore how discrepancies between the model's predictions and experimental data may reflect signatures of transcriptional activity, chromatin openness, and genomic fragility.

We begin our analysis with a fundamental inquiry: how accurately can a kinetic model of replication predict genome-wide timing? Our model acts as a null hypothesis, representing how replication should occur in the absence of perturbation from genomic features. The central aim is to identify loci where the model's predictions diverge from experimental observations, highlighting regions that may ex-

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

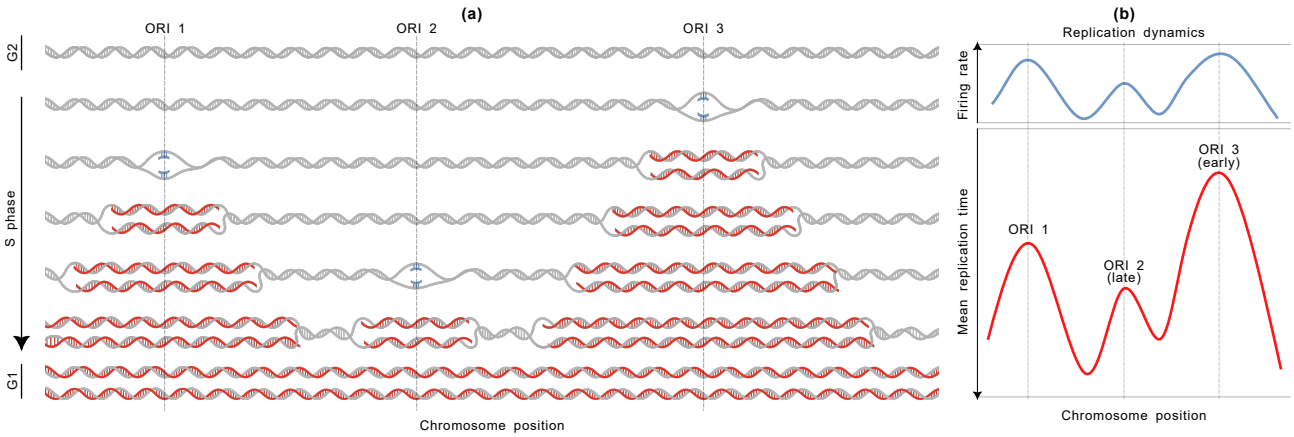


Figure 1. A kinetic model of DNA replication.
(a) Replication initiates at specific origins that are fully licensed by the end of G1 phase. During S phase, replication forks progress bidirectionally from these origins, passively replicating the DNA until they merge with forks from adjacent origins or reach the ends of chromosomes, thereby completing replication and entering G2. In this example, three origins (ORI 1, ORI 2, and ORI 3) fire at different times, with nascent DNA strands shown in red. At the end of replication, two identical copies of the original DNA template are formed. **(b)** The figure illustrates the expected inverse but non-trivial correlation between firing rates (top) and replication timing (bottom, with an inverted y-axis). In a model where the firing time of each origin is an exponentially distributed random variable, the firing rate is the parameter of this distribution and tends to decrease as replication timing increases, indicating that regions with higher firing rates replicate earlier in S phase. Replication timing, measured by Repli-seq, shows the average replication time across a population of cells, with peaks in the profile corresponding to potential replication origins. ORI 2 is positioned in a late-replicating region, while ORI 3 replicates earlier, as indicated by their relative positions on the timing curve. Adapted from Hulke et al. (2020).

perience replication stress or other anomalies. By deriving a closed formula for the expected time of replication at each genomic site, we establish a solid mathematical framework to support our computational simulations.

Our workflow is simple: using only timing data as input, along with minimal genomic parameters such as potential origin locations, the model determines firing rates and predicts timing profiles plus other key kinetic features like fork directionality and inter-origin distances. Researchers with replication timing data can use this model to rapidly generate precise replication dynamics profiles without extensive computational expertise, revealing factors that influence replication timing and genome instability across various contexts.

Despite significant advances in mathematical modelling (Jun et al., 2005; Jun and Bechhoefer, 2005; de Moura et al., 2010; Retkute et al., 2012), deriving a position-specific, data-fitted model that precisely links replication timing to origin firing has remained a challenge. While some approaches rely on neural networks to infer probabilistic landscapes of origin efficiency (Arbona et al., 2023), ours differs by deriving a closed-form relationship between timing and firing. Rather than relying on complex inference techniques, our model abstracts intrinsic firing rates without directly tying them to specific biological mechanisms such as licensing or activation. This allows a precise fit to observed timing data and enables simulation of genome-wide dynamics in a direct and interpretable manner. Our approach improves existing fitting methods by adopting a convolution-based interpretation of the timing programme. Using process algebras from concurrency theory (Boemo et al., 2020), we model replication forks and origins as interconnected entities, simulating their behaviour across the genome. The key contribution of this work is demonstrating how a theoretical description of replication timing helps uncover links between timing, genomic stability, and other essential genomic processes.

Methods

Modelling assumptions

We aim to identify and quantify regions of the genome where replication timing deviates from model predictions, hereafter referred to as replication timing misfits, which may indicate potential sites of replication stress or instability. To accomplish this, we model the complex, nonlinear relationship between origin firing rates and replication timing (Figure 1b) and fit these rates to experimental timing data. This approach enables investigation using replication forks, origins, and DNA templates as the level of abstraction. In particular, we do not differentiate between leading and lagging strands, as the formation and joining of Okazaki fragments are not explicitly included in the model. By focusing on the fundamental kinetics driving replication, we gain a clearer understanding of how firing influences outcomes.

Our model operates under several key assumptions. The time an origin fires is modelled as an exponentially distributed variable (independent of fork movement and firing of other origins), and fork movement as an exponentially distributed random variable (independent of origin firing and movement of other forks). We also assume a constant rate of fork movement throughout (no fork stalling at obstacles); then, forks advance smoothly until encountering another fork or a chromosome end. This assumption avoids overfitting and indirectly emphasizes the role of origin firing. The key variable is the origin firing rate. This encompasses origin licensing and activation, plus contributions of all other proteins and pathways within this process. While a strong assumption, it is justified by the fact that firing rates effectively capture the collective outcome of all these underlying processes without explicitly representing molecular detail. This makes the model both tractable and capable of producing accurate genome-wide predic-

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tions. We further sub-divide the genome into 1 kb intervals (sites), and assign to each a non-zero firing rate determined by a governing equation that links timing with firing. This resolution offers a balance between computational efficiency and biological realism. Although any site is a potential origin, passive replication and low firing rates ensure the expected sparsity of origins seen in the genome.

Mathematical modelling of replication

Consider a DNA molecule with n discrete genomic loci, where each locus can potentially act as an origin that fires at rate f to initiate a fork that progresses bidirectionally with speed v, typically measured in kilobases per minute (kb/min). We aim to determine the average time required for a site to either initiate replication, or to be passively replicated by an approaching fork (i.e., its expected replication time). Initially, we assume that all origins fire at the same rate, f, but later relax this assumption to allow for variations in firing rates across different origins. In addition, by considering a sufficiently large chromosome, we ensure that the effects of chromosomal ends are negligible. Nonetheless, the framework can easily be extended to account for such effects, though they are not critical for the broader analysis.

Expected time of replication. Let T be the time a site takes to fire or be passively replicated by a fork. We assume initially that all origins fire at the same rate, f. One may think of T as an explicit function of origin firing times A_i, where $A_i \overset{\text{iid}}{\sim} \text{Exp}(f)$. In particular, $\mathbb{E}[A_i] = 1/f$. We index each site by its distance from the origin of interest, given by $|i|$. Notice that $i = 0$ corresponds to the focal origin, and v is interpreted as the number of replicated sites per time unit. We have

$$T = \min_i \{A_i + |i|/v\} \tag{1}$$

since it takes time $|i|/v$ for a fork initiated at site i to reach the origin of interest. Next, we compute the cumulative distribution function. The minimum in Eq. (1) is greater than some t if all its components are, which occurs with probability

$$P(T > t) = \prod_i \min\{1, \exp(-f(t - |i|/v))\} \tag{2}$$

since $A_i > 0$ and $A_i \overset{\text{iid}}{\sim} \text{Exp}(f)$. Hence, the expectation of the replication time for any one site is given by

$$\mathbb{E}[T; n] = \int_0^\infty \prod_i \min\{1, \exp(-f(t - |i|/v))\}\, dt \tag{3}$$

where the product is taken over all n sites. This integral can be partitioned across each interval for which $|i| \le vt \le |i+1|$. Within these intervals, the integrands adopt the form ae^{-bt}, thereby permitting analytical evaluation. In the general case, the result depends on the parity of n. See Supplementary Note (SN) 1.1 for an expression of $\mathbb{E}[T; n]$.

As $n \to \infty$, a general expression of the expected replication time for each origin can be written as

$$\mathbb{E}[T; \infty] \equiv \frac{1}{f} \sum_{k=0}^{\infty} \frac{e^{-fk^2/v} - e^{-f(k+1)^2/v}}{2k+1}. \tag{4}$$

With $v = 1.4$ kb/min (Conti et al., 2007), Figure 2a shows the dynamics of $\mathbb{E}[T; n]$ for increasing values of n. By relating Eq. (4) to the family of theta and Dawson functions (Tyurin, 2002; Temme, 2010), the following approximation holds (see SN1.2 for a detailed proof)

$$\mathbb{E}[T; \infty] \simeq \frac{1}{2}\sqrt{\frac{\pi}{fv}}. \tag{5}$$

Provided replication timing data $\{T_i\}_{1 \le i \le n}$, we have the following inversion

$$f_i \simeq \frac{\pi}{4v} T_i^{-2} \tag{6}$$

which provides a first estimate for the intrinsic firing rate of an origin, given its time of replication. Note that Eq. (6) is an approximation under the specific assumption that firing rates are uniformly constant across the genome, a simplification that, intriguingly, offers a reasonably accurate initial estimate for the firing rate distribution in most instances. The fidelity of this approximation is closely tied to fork speed v and the average of the timing dataset, topics that will be elaborated subsequently.

A generalisation. Experimental data support the idea that different origins fire at different rates (Leonard and Méchali, 2013; Lu and Pickett, 2022). While our introductory argument assumes a constant firing rate f across the genome, we should, in general, expect $A_i \sim \text{Exp}(f_i)$. Then, the replication time definition in Eq. (1) should include the site-specific indexation, for $1 \le j \le n$, as follows

$$T_j = \min_i \{A_i + |i - j|/v\} \tag{7}$$

with indexes congruent modulo n, that is, $|i - j| \in \mathbb{Z}/n\mathbb{Z}$ (see SN1). Following a similar argument, the general expression for $\mathbb{E}[T_j; \infty]$, with general firing rates $\{f_i\}$, is approximately given by

$$\sum_{k=0}^{\infty} \frac{e^{-\sum_{|i| \le k}(k - |i|) f_{j+i}/v} - e^{-\sum_{|i| \le k}(k+1-|i|) f_{j+i}/v}}{\sum_{|i| \le k} f_{j+i}}. \tag{8}$$

When $f_j = f$, $\forall j$, Eq. (8) is reduced to Eq. (4). While Eq. (8) holds true for an infinitely large genome, in practical terms this series can be limited to $0 \le k \le R < n/2$, for some large enough R. This parameter represents the radius of replication influence: the distance within which neighbouring origins $\{j - R, ..., j - 1, j + 1, ..., j + R\}$ are assumed to affect the timing of a focal origin j. In other words, while every firing origin does theoretically affect replication timing at any other location, this effect decays rapidly with distance from the origin of interest j. Numerically, the finite version of Eq. (8) should mimic the average

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

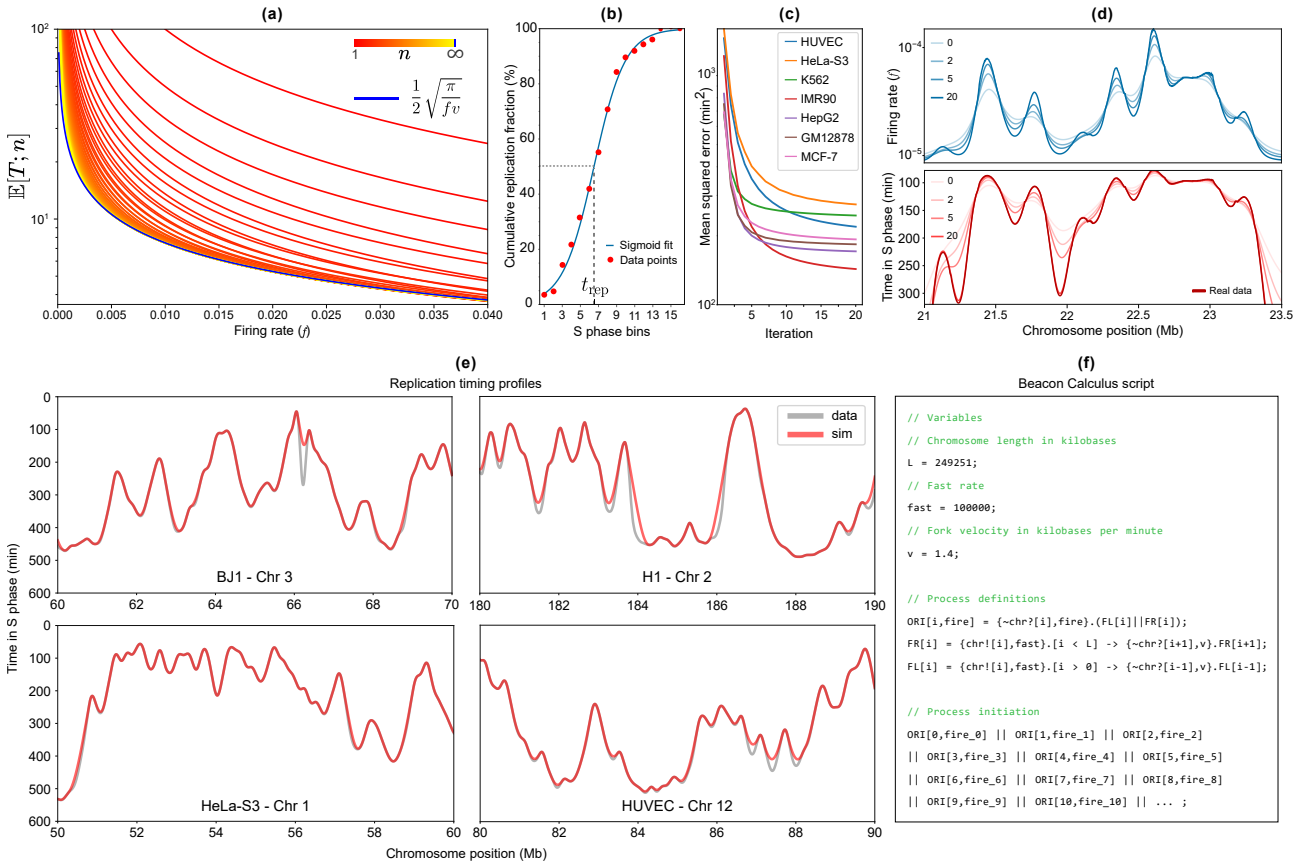


Figure 2. Fitting the model.
(a) Replication asymptotics under uniform firing: log plot of expected time of replication, $\mathbb{E}[T;n]$, as a function of f and n, for $1 \leq n < \infty$, with $v = 1.4$ kb/min. As $n \to \infty$, $\mathbb{E}[T;n]$ approximates an inverse power law (blue). **(b)** Curve fitting for cumulative replication in S phase. Red markers depict example data points from a high resolution Repli-seq heatmap that shows the cumulative percentage of completed replication across 16 S phase bins. The blue line is the curve fitted to this data, while the dashed grey line indicates t_{rep}, the point in S phase when 50% of replication is achieved across the cell population. **(c)** Whole-genome mean-squared error between simulated timing profiles and real data for 7 cell lines, in min^2. Fitting each line took $\sim$ 3 minutes on a HPC platform (one CPU). **(d)** Progression of the fitting algorithm over 20 iterations for chromosome 2 in the BJ line on firing rates (above), with iteration 0 corresponding to the initial inverse power law estimate, given by Eq. (6), and the corresponding timing profile (below). **(e)** Observed (Repli-seq) timing against the simulated (bcs) profiles for different lines and genomic regions. **(f)** Model written in the Beacon Calculus. Origin firing processes take their location, `i` (1-kb resolution), and firing rate `fire`, as parameters, triggering two replication fork processes, `FL` (left-moving) and `FR` (right-moving). Replication terminates when all locations have been replicated. The simulation begins by invoking the `ORI` processes, where `fire_i` corresponds to the firing rate values for each origin `i`, as determined by fitting Eq. (8).

replication timing obtained from computational simulations and it will be crucial in solving the fitting problem efficiently. Ideally, we would like to compute the rates $\{f_i\}$ as a function of the expected values of T_j. Our goal is then to find a solution to Eq. (8), given data on $\{\mathbb{E}[T_j;n]\}$, for large n. Alternative frameworks inspired by the analogy between DNA replication and crystal growth have been explored by Kolmogorov (1937); Jun et al. (2005); Jun and Bechhoefer (2005); Jun and Rhind (2008). Such work reveals other relevant replication metrics, such as inter-origin distances (Herrick et al., 2002). Our formulation extends these approaches by giving an estimation of origin firing rates from discrete replication timing data across the entire human genome, which is discussed next.

Replication timing data

Replication timing data was sourced and processed from two key databases: Encyclopedia of DNA Elements (ENCODE) (Hansen et al., 2010; Davis et al., 2018) and high-resolution Repli-seq from Zhao et al. (2020). To ensure data consistency and reliability, extensive filtering, and scaling steps were performed on all data sets. We analyse data from: HUVEC (human umbilical vein endothelial cells), HeLa-S3 (clonal derivative of the parent HeLa, an immortalised cervical cancer line), BJ (normal skin fibroblast), IMR90 (lung fibroblast), K562 (lymphoblast cells), GM12878 (lymphoblastoid line), HepG2 (hepatocellular carcinoma line), MCF-7 (breast cancer line), HCT (colorectal carcinoma line), plus H1 and H9 (embryonic stem cell lines). Data for HUVEC, HeLa, BJ, IMR90, K562, GM12878, HepG2, and MCF-7 cells was sourced from the ENCODE database, on the GRCh37 (hg19) human genome assembly (Hansen et al., 2010; Davis et al., 2018). In contrast, data for HCT, H1, and H9 was obtained from high-resolution Repli-seq, on the GRCh38 (hg38) assembly (Zhao et al., 2020; Wang et al., 2021).

Regarding ENCODE Repli-seq, timing data from each cell line were analysed across 6 cell cycle fractions: G1/G1b, S1, S2, S3, S4, and G2, given as a wavelet-smoothed signal to generate a continuous portrayal of replication across the genome (Thurman et al., 2007). Importantly, we rescaled the original wavelet signal, initially normalised from 0 to 100, by a factor of 6 to better align with an approximately 8-hour S phase.

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Following standard Repli-seq methods, we applied a sigmoidal fit to the cumulative replication fraction, F_{rep}, to determine replication timing according to Zhao et al. (2020). We consider the median replication time, t_{rep}, defined as the bin value t where $F_{\text{rep}}(t) = 50\%$, indicating that half of the cell population has completed replication (Figure 2b). Although Eq. (8) theoretically represents the mean replication timing, it aligns closely with the median observed in Repli-seq data, as replication timing distributions generally exhibit a near-symmetric sigmoidal pattern. Additionally, the median is more robust to experimental noise and outliers, making it a practical and reliable measure in high-throughput experiments. Although recent studies have determined telomere timing data (Massey and Koren, 2022), we do not incorporate this into our analysis.

Repli-seq data shows consistent patterns across different cell lines. We present representative results from multiple lines (Figure 2e), but specific analyses may be more suitable for certain ones, depending on the availability and quality of the data. Although regions with repetitive sequences or low complexity are often mapped poorly using Repli-seq data (Hansen et al., 2010; Zhao et al., 2020), these regions account for $\sim$ 20% of the genome and show only a weak correlation with high-misfit regions (phi coefficient = 0.21). Therefore, we retain this data in our analysis, as its impact is minimal (see SN2.3).

Fitting algorithm

We develop an algorithm to efficiently fit genome-wide replication timing data, processing over 3,200,000 potential origins per genome by leveraging the mean-field dynamics captured by Eq. (6). While every site is treated as a potential origin, the algorithm effectively suppresses many by assigning them negligible firing rates, reflecting the selective activation of origins in the genome. For large n, Eq. (8) provides an excellent estimate, allowing us to apply a fitting algorithm directly to the theoretical expectation, rather than relying on averaged outputs from multiple simulations. The radius of neighbouring influence, R, may be refined for optimisation. At each site j, we set

$$f_j(0) = \frac{\pi}{4v} T_j^{-2} \tag{9}$$

$$f_j(k+1) = f_j(k) \left(\frac{\tilde{T}_j(k)}{T_j} \right)^{\alpha} \tag{10}$$

where $\tilde{T}_j(k)$ is the average replication time at iteration k, via Eq. (8), and T_j is the real data value. The parameter α is adjusted to guarantee convergence towards a stable firing rate distribution. The extent of each misfit is measured by the squared difference (fit error) between observed and expected replication timing, in min^2, at each site. Numerical results and Eq. (6) suggest $\alpha = 2$ is a reasonable compromise between error minimisation and speed. The optimised algorithm considers a convolution interpretation of Eq. (8) (SN2.2). Remarkably, the firing rates for each cell are fitted in an average time of approximately 3 minutes using 1 CPU on a high-performance computing platform equipped with Intel Ice Lake architecture (Figures 2c-e).

Simulations

To simulate replication, we use Beacon Calculus (`bcs`), a process algebra designed for simulating biological systems (Boemo et al., 2020). Within this framework, replication is modelled using three core processes: replication origins (`ORI`), left-moving forks (`FL`), and right-moving forks (`FR`). Each process is associated with a specific position on the chromosome of length `L`, and origins have an additional parameter, the firing rate, `fire`, or f in our model (Figure 2f).

In `bcs`, v is understood as the rate of replication by a moving fork, which is held constant. This differs from the constant fork speed assumption underlying Eq. (8). Specifically, in the `bcs` case, the time F_k required for a fork to replicate k consecutive sites follows an Erlang(k, v) distribution, meaning that $\mathbb{E}[F_{|i-j|}] = |i-j|/v$, which mirrors the approximation used in Eq. (7). Therefore, when averaged over a sufficiently large number of simulations, the stochastic deviations in numerical simulations become negligible and they do not compromise the broader analysis or conclusions.

To track the progress of replication, the model marks regions of the chromosome that have been replicated, allowing us to monitor replication dynamics accurately. In all `bcs` simulations, fork speed was set to 1.4 kb/min (Conti et al., 2007), and results were averaged over 500 simulations, with the radius of influence set to $R = 2000$ kb, as previously defined.

Results

Predicting genome-wide replication

After assigning the time of replication (determined using Repli-seq data) to every 1 kb segment of the genome in 11 different human cell lines, site-specific firing rates are fit to the data via Eq. (8). Then, replication is simulated using Beacon Calculus (`bcs`), a concise process algebra ideal for concurrent systems. We then explore patterns of replication seen after averaging 500 simulations for each of 11 different cell lines (Figure 3a).

We begin by comparing the experimental timing profile with the one obtained from Eq. (8). Note that this is equivalent to averaging the timing profiles from a large number of `bcs` simulations, which also allows us to save significant computational resources when computing timing alone. An example for chromosome 1 in HUVECs is shown in Figure 3b.i. As expected, some regions replicate early (e.g., around 173 Mb) and others late (e.g., around 171 Mb). There is generally excellent concordance between the model's output compared with the experimental input.

Our focus is on regions with high misfit error (shaded yellow and red areas; Figure 3b.i), where concordance breaks down because Eq. (8) predicts an earlier replication time than observed experimentally. Instances where it predicts later replication compared to data are exceedingly rare, underscoring the algorithm's reliance on higher firing rates to achieve the best fit with theoretical expectations.

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

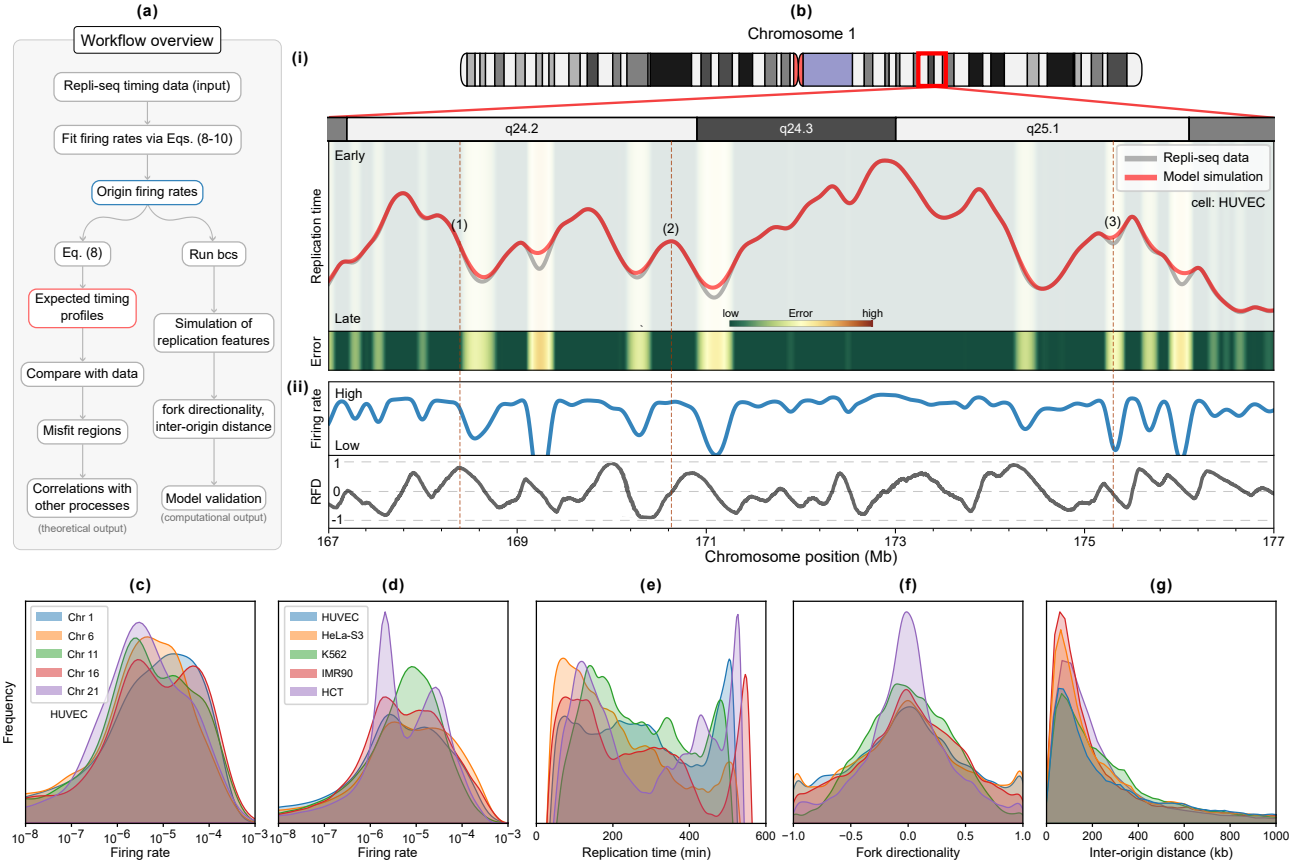


Figure 3. Predicting genome-wide features of replication.
(a) Overview of the main model and analysis. Starting with Repli-seq timing data, origin firing rates are fitted through Eqs. (8-10). These rates generate expected timing profiles for comparison with experimental data to identify regions of timing misfits and fork stalling, which are analysed for correlations with other genomic processes. Simulations of replication features, such as fork directionality and inter-origin distances, validate the model against the literature. **(b)** Example of main modelling outputs from a region in HUVECs. **(i)** Replication timing of both experimental and simulated data, and the magnitude of the misfit (error) for replication timing obtained with HUVECs in a region where replication forks often stall; this leads to elevated errors that the model struggles to capture accurately. **(ii)** Inferred origin firing rates, and fork directionality, scaled between -1 (leftward) and 1 (rightward). We highlight three regions of interest: (1) A passively replicated site predominantly replicated by rightward-moving forks (RFD $\sim$ 1); (2) A likely origin, characterised by a high firing rate and an RFD of 0; (3) A poorly fitted region between two origins with a low firing rate determined by the fitting algorithm with RFD of 0 (an equal likelihood of replication by leftward- and rightward-moving forks). **(c)** Kernel density estimate (KDE) of firing rate distributions across selected chromosomes in HUVECs. **(d-g)** KDEs comparing genome-wide features—including firing rates, replication timing, fork directionality, and inter-origin distances—across different cell lines. All distributions align with experimental observations. Areas under curves are equal to 1, while y-axis numerical values are omitted to emphasize relative shapes and distributions rather than absolute magnitudes.

While firing rates are directly inferred from Eq. (8), replication fork directionality (RFD) is calculated as the proportion of cell cycles (or bcs simulations) in which a given site is replicated by rightward versus leftward forks. RFD values range from -1 (always replicated by leftward forks) to 1 (always replicated by rightward forks), with intermediate values indicating a mix of replication directions across simulations (Figure 3b.ii).

To validate the model, we examine global distributions of multiple features. Despite little variation in firing in HUVECs (Figure 3c), HCT exhibits a pronounced bimodal pattern, likely driven by differences in data sources (Figures 3d-e; Zhao et al. (2020)), which may affect how replication timing and origin firing rates are captured. Regarding RFD, our results demonstrate a balanced bidirectional fork movement, with fork directionality symmetrically distributed and accumulating around zero, indicating efficient replication progression (Figure 3f). This pattern aligns with recent quantifications of fork directionality in human cells (Anderson et al., 2024). While determining inter-origin distances (IOD) is straightforward from our simulations, doing so from DNA-fiber experiments remains challenging due to technical limitations and potential biases (Técher et al., 2013; Quinet et al., 2017). Nevertheless, simulations show a concentration of IODs within the commonly observed range of 100–200 kb (Figure 3g; Conti et al. (2007)).

Although these results validate the model against established metrics, its broader ability to simulate other features, like replicon lengths and active fork numbers, underscores its value in capturing the full spectrum of replication dynamics. The most compelling insight, however, comes from examining regions where the model's predictions diverge from data, as these discrepancies may coincide with critical sites of genomic instability, revealing areas of unique biological interest, which we address next.

Hotspots of instability

We now determine genome-wide error profiles in all 11 cell types (Figure 4a illustrates that for chromosome 1). Remarkably, some of the regions that fit poorly are found in all cell lines (despite using different genome builds); this underscores the robustness of profiles across cell types

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

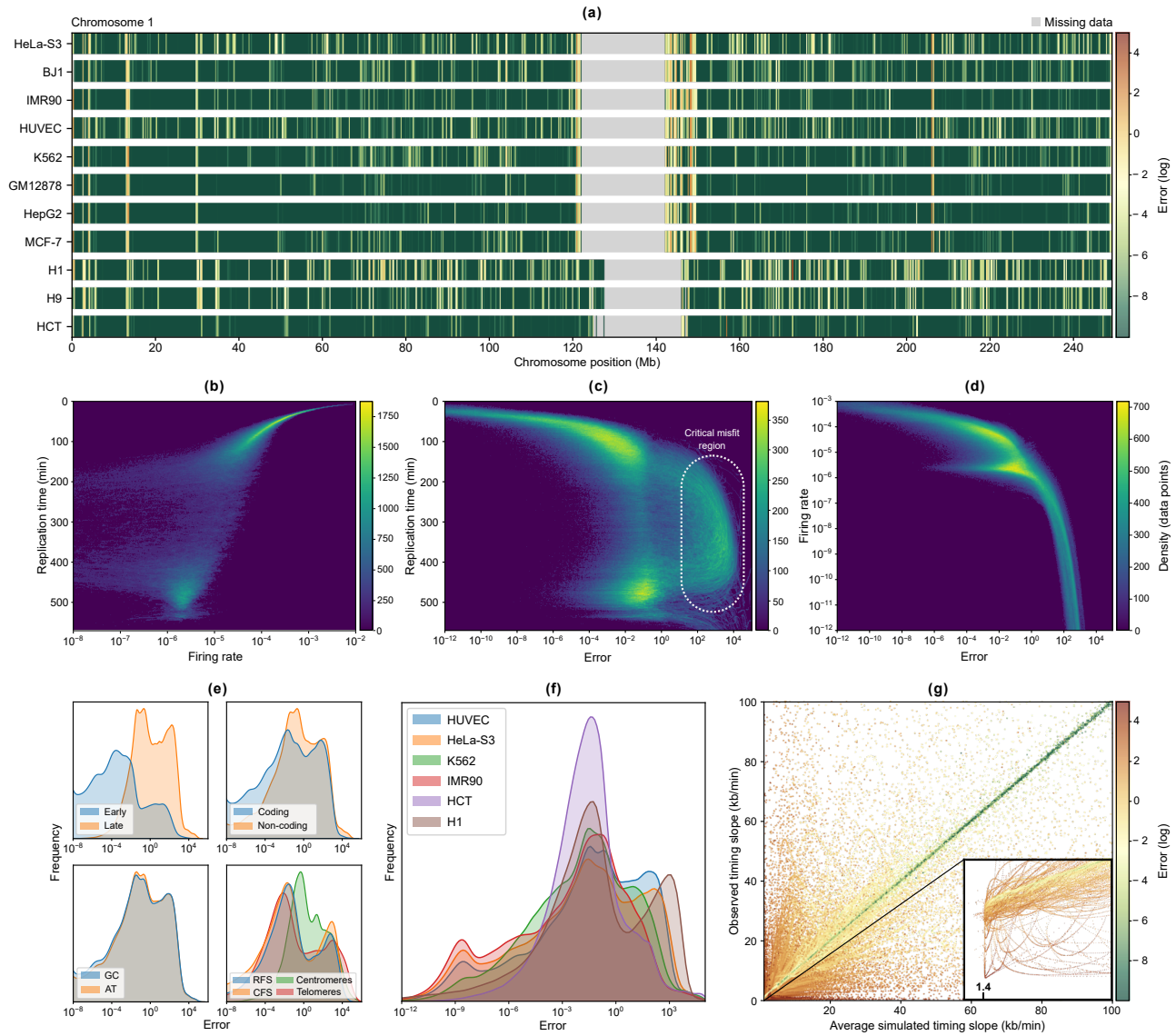


Figure 4. Detecting discrepancies in replication timing determined experimentally and in simulations.
(a) Normalised error plots (red: high error, green: low error) highlighting deviations between simulated and experimental replication timings (chromosome 1 in various human cell lines). Grey areas: missing or unavailable data. Chromosome positions are based on the hg19 build, except for HCT, H1, and H9 that used the hg38 build. (**b-d**) Density scatter plots illustrating key relationships in H1 cells (averages of 500 simulations). We look at the pairwise combinations of three variables: replication time, firing rate, and error. Density is represented as the number of data points per pixel on an 80 DPI (dots per inch) resolution across all plots. In **(b)**, the inverse correlation between replication timing and firing rate is evident, with greater variability in firing rates observed during late S phase. **(c)** highlights the relationship between replication timing and error, showing that high errors are distributed throughout S phase (dotted oval). **(d)** illustrates the branching relationship between firing rate and error. **(e)** Error distributions in HUVEC cells. Early-replicating regions tend to be error prone, coding and non-coding regions have broadly similar patterns, AT-rich regions are much like GC-rich ones, and different genomic regions have characteristic profiles. **(f)** Genome-wide error profiles in different cells. **(g)** Scatter plot comparing average simulated timing slope, indicative of the rate of replication progress over time, against observed data, color-coded by associated error. The zoomed-in region at $[1.2, 2] \times [0, 2]$ kb/min highlights the 1.4 kb/min lower bound on the simulated slope. Each dot represents a simulated-observed data pair, with the strand-like continuity resulting from the high resolution of our 1 kb model.

(Bracci et al., 2023; Müller and Nieduszynski, 2012). Replication timing and firing rates are strongly negatively correlated (Spearman's rank correlation of $\sim$ -0.89; Figure 4b); regions with higher firing rates tend to replicate earlier. Late-replicating regions also have a wide spread of low firing rates, reflecting a pattern captured by the fitting algorithm. Additionally, the lowest errors are seen in the earliest replicating regions, moderate ones in both early- and late-replicating regions, and the highest are distributed throughout the mid-to-late S phase (Figure 4c). This suggests misfits increase as S phase progresses and fewer firing events occur. Low firing rates are also associated with high errors (Figure 4d; note the branched profile, reflecting difficulties in accurately modelling high-to-low firing rate transitions).

Timing misfits are predominantly concentrated in late-replicating regions (Figure 4e). This is consistent with prior results suggesting that the replication machinery encounters more obstacles towards the end of S phase (Branzei and Foiani, 2010; Colicino-Murbach et al., 2024). Additionally, errors exceeding 10^4 (min^2) are more frequent in non-coding regions compared to coding ones, indicating a potential vulnerability of non-coding DNA to replication stress. Misfits also vary between cell lines, with HCT displaying a distinct pattern likely due to differences in data processing (Figure 4f). Similar disparities were observed previously (Figure 3d), hinting at the potential for cell line-

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

specific analyses to offer further insights. However, given our focus here, we leave a detailed analysis of these dynamics for future exploration.

In regions with infrequent origin firing, the slope of the timing curve—representing the rate of replication progress over time—is primarily governed by fork speed, establishing an effective lower bound of 1.4 kb/min (Figure 4g). This constraint becomes most evident in regions where observed slopes fall below such bound, resulting in error accumulation around slower-replicating areas. Origin competition, where nearby origins fire at similar times, further compounds these errors, producing timing 'valleys' between origin firing peaks. These patterns highlight regions of potential stress, suggesting areas for further study.

Fragile sites and large genes

Fragile sites are cytogenetically defined gaps and breaks in metaphase chromosomes (Li and Wu, 2020; Glover et al., 1984); examples include FRA3B (Le Beau et al., 1998) and FRA16D (Palakodeti et al., 2004). They are often seen after partially inhibiting DNA synthesis or applying other replicational stresses (Durkin and Glover, 2007; Glover et al., 2017). They also contain few origins (Sinai and Kerem, 2018), and probably arise due to fork stalling or collapse (Kaushal and Freudenreich, 2019; Franchitto and Pichierri, 2014). Fragile sites can be broadly categorised into common fragile sites (CFSs) present in the whole population and rarer ones (RFSs) found in only a few individuals (Schwartz et al., 2006). We now consider both classes, using site locations obtained from the HumCFS database (Kumar et al., 2019) and gene locations from the GENCODE Genes track v46 (Frankish et al., 2023).

As seen in Figure 4c, replication timing misfits are most pronounced in mid-to-late S phase, where our model struggles to capture timings accurately. Regions such as centromeres and telomeres, as well as most fragile sites often map to regions of high error, particularly during late S phase (Figures 5a-d). FRA3B and FRA16D show even higher median misfit lengths, suggesting these regions are especially challenging for the model to fit (Figure 5e). Similarly, large genes in fragile sites, such as *CNTNAP2*, *LRP1B*, and *FHIT*, also exhibit substantial error (Figures 5f-g). Under a certain error threshold, large genes overlapping with misfit regions, including those in fragile sites, may be easily identified through our model (Figure 5h). Approximately 30% of these genes overlap with fragile sites, reinforcing the established link between late replication timing and the structural demands of large gene transcription. Notably, chromosomes 15, 20, and 22 show no significant misfits associated with fragile sites, likely due to a lower abundance or reduced activity of fragile sites on these chromosomes.

While not all fragile sites follow this pattern, the observed timing misfits at large genes suggest broader genomic regulation factors that may underlie replication stress. This analysis motivates further investigation into whether transcriptional activity and chromatin state are linked to these misfits globally.

Transcription and chromatin state

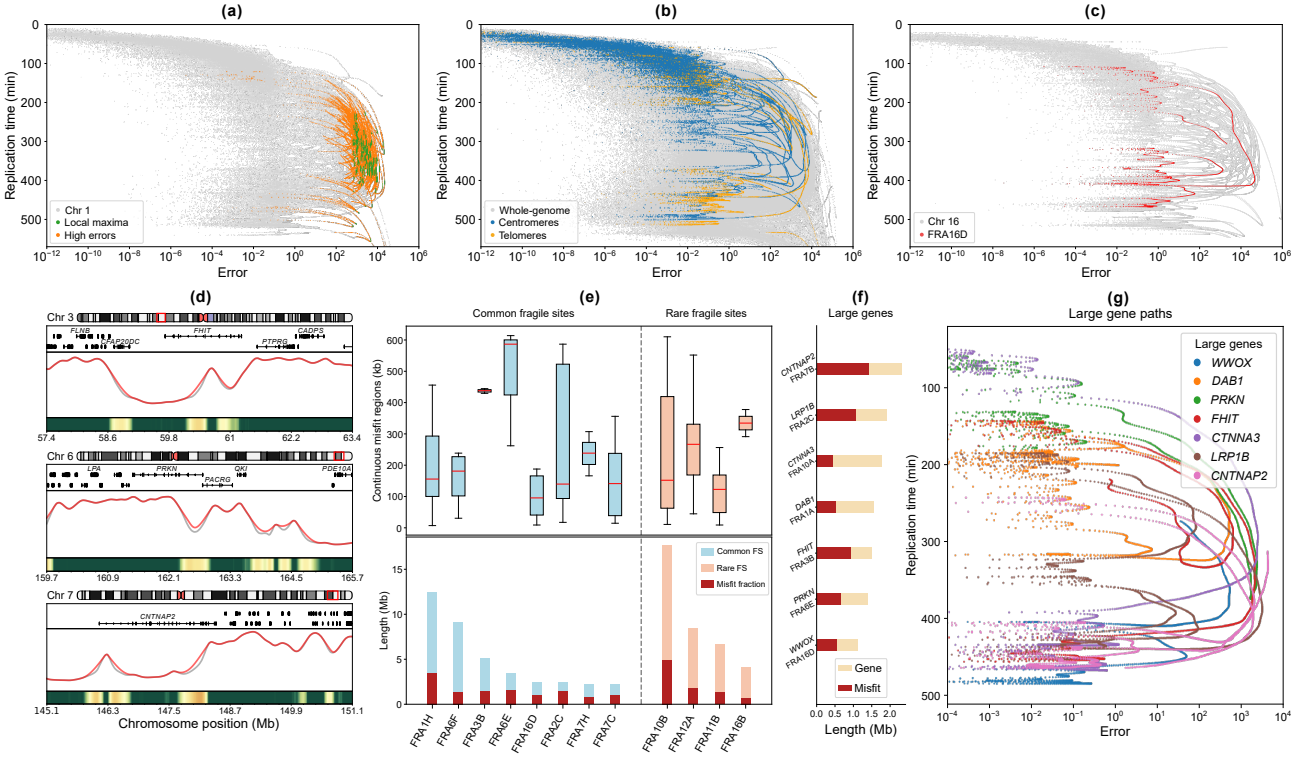
Transcription and replication have long been recognised to interact in complex and sometimes conflicting ways, particularly in regions like fragile sites (Knott et al., 2009). Previous studies have shown that transcription-dependent barriers can obstruct replication fork progression, leading to stalling or collapse, while large genes associated with CFSs often lack sufficient replication initiation events, forcing replication forks to traverse longer distances from adjacent origins and further delaying replication completion. (Blin et al., 2019). This delay is particularly pronounced in transcriptionally active regions, where the overlap between transcription and replication processes heightens the risk of replication stress. However, this is not always the case, as chromatin structure can play a more dominant role in dictating replication timing discrepancies.

Building on the previous results, we now turn our attention to the interaction between transcription, chromatin structure, and replication dynamics. Regulatory elements like transcription promoters and enhancers are marked by histone modifications such as H3K4me3, indicating active transcriptional regions (Huang et al., 2019). These open chromatin regions can be identified through DNase I hypersensitivity (DHS) and ChIP-seq assays (Cockerill, 2011; Young et al., 2011). By integrating data from ChIP-seq, RNA-seq, and GRO-seq (Cockerill, 2011; Crawford et al., 2004; Marguerat and Bähler, 2010; Lopes et al., 2017), we can assess how chromatin accessibility, histone modifications, and transcriptional activity are associated with replication timing and other DNA-dependent processes.

Comparing replication dynamics with transcriptional activity reveals that regions with high GRO-seq signals, indicative of active transcription, align with peaks in H3K4me3 and DNase I hypersensitivity signals, and tend to exhibit lower replication timing errors and higher firing rates (Figures 6a). Spearman rank correlation analyses reveal varying degrees of association between variables (Figure 6b). This method was chosen due to its suitability for non-normally distributed data and its ability to capture monotonic relationships, reflecting the ranked nature of our genomic features. Pearson and Kendall's Tau tests were conducted for comparison (see SN2.4). The consistently higher Spearman rank correlations indicate a strong monotonic relationship, particularly between DNase I hypersensitive sites and firing rates, as well as between promoters and firing rates. This suggests that chromatin state, indicated by DNase I sensitivity and promoter activity, is key in promoting origin firing. The strength of these correlations underlines how chromatin accessibility facilitates replication initiation, even amid non-linear interactions.

We observed a moderate to strong negative correlation between GRO-seq and replication misfits across all cell lines. This suggests that the replication machinery may encounter fewer impediments in regions with active transcription. A possible explanation is that transcriptionally active regions are more likely to be in an open chromatin state, reducing the mechanical barriers to replication fork progression and lowering the chances of replication stress. More-

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



(h) Misfit genes

CFS RFS Both

Chr										
1	*AGBL4*	*KAZN*	*NEGR1*	*RABGAP1L*	*RYR2*	*DNM3*	*ST6GALNAC3*	*KCNH1*	*HMCN1*	*PLD5*
2	*LRP1B*	*DPP10*	*NRXN1*	*THSD7B*	*NCKAP5*	*CNTNAP5*	*ALK*	*AFF3*	*MYT1L*	*KCNS3*
3	*FHIT*	*RBMS3*	*TBC1D5*	*ROBO1*	*LSAMP*	*CADM2*	*CACNA2D3*	*EPHA6*	*ZBTB20*	*LPP*
4	*FSTL5*	*LRBA*	*CFAP299*	*RASGEF1B*	*ANK2*	*TENM3*	*SORCS2*	*STK32B*	*MAML3*	*AFG2A*
5	*PDE4D*	*TENM2*	*CDH18*	*SGCD*	*SLIT3*	*SPOCK1*	*FER*	*EDIL3*	*HCN1*	*FBXL7*
6	*PRKN*	*NKAIN2*	*GRIK2*	*ADGRB3*	*GMDS*	*FARS2*	*PKHD1*	*TRDN*	*SLC35F1*	*ZDHHC14*
7	*CNTNAP2*	*MAGI2*	*DPP6*	*SDK1*	*IMMP2L*	*DGKB*	*SUGCT*	*BBS9*	*CDK14*	*ELMO1*
8	*NRG1*	*VPS13B*	*NKAIN3*	*UNC5D*	*XKR4*	*EXT1*	*MCPH1*	*ASPH*	*EBF2*	*C8orf44-SGK3*
9	*PTPRD*	*LINGO2*	*ADAMTSL1*	*TRPM3*	*DENND1A*	*BNC2*	*ROR2*	*NFIB*	*RFX3*	*SLC24A2*
10	*PCDH15*	*NRG3*	*KCNMA1*	*GRID1*	*PARD3*	*ANK3*	*SORCS1*	*PLXDC2*	*CACNB2*	*ABLIM1*
11	*DLG2*	*LRRC4C*	*CNTN5*	*NELL1*	*TENM4*	*NAV2*	*KIRREL3*	*GRM5*	*SOX6*	*DCDC1*
12	*ANKS1B*	*MGAT4C*	*TMTC2*	*ANO4*	*TRHDE*	*CNTN1*	*SLC2A13*	*ABTB3*	*PTPRO*	*SLCO1B3-B7*
13	*NBEA*	*MYO16*	*KLHL1*	*DCLK1*	*CLYBL*	*FREM2*	*CLDN10*	*CAB39L*	*SCEL*	*TNFRSF19*
14	*RAD51B*	*GPHN*	*TTC6*	*TSHR*	*TRAF3*	*CDC42BPB*	*BAZ1A*	*MIA2*	*LIN52*	*SOS2*
15	*UNC13C*	*FMN1*	*ADAMTS17*	*IGF1R*	*APBA2*	*LRRC49*	*RNF111*	*RFX7*	*SHC4*	*TRPM7*
16	*WWOX*	*RBFOX1*	*CDH13*	*GSE1*	*FTO*	*ZNF423*	*ITFG1*	*ACSM3*	*ADAMTS18*	*CFDP1*
17	*ASIC2*	*SHISA6*	*ACACA*	*SPECC1*	*ARSG*	*VMP1*	*MAP2K4*	*SMURF2*	*TADA2A*	*DHRS7B*
18	*DCC*	*DLGAP1*	*CCDC178*	*L3MBTL4*	*LDLRAD4*	*KIAA1328*	*NEDD4L*	*LOXHD1*	*MAPK4*	*CCDC102B*
19	*MARK4*	*INSR*	*MUC16*	*TDRD12*	*ZNF83*	*URI1*	*ZNF569*	*NLRP11*	*AP1M1*	*NLRP4*
20	*PTPRT*	*PLCB1*	*PLCB4*	*PAK5*	*EYA2*	*RIN2*	*SYNDIG1*	*NCOA3*	*ZFP64*	*RALGAPB*
21	*CHODL*	*GET1- SH3BGR*	*TTC3*	*SH3BGR*						
22	*BCR*	*DGCR2*	*GAB4*	*YPEL1*	*PPIL2*	*MAPK8IP2*	*ARSA*			

Gene length

Figure 5. Timing errors in fragile sites and large genes

(a) Replication timing vs. error on chromosome 1 in H1, highlighting regions with local maxima in error and neighbouring high-error zones (within a 300 kb radius). The threshold for identifying local maxima in errors is set at $10^{2.8}$ (min^2). Each dot represents an error-timing data pair, with the strand-like continuity arising from the high-resolution of our 1 kb model. **(b)** Genome-wide scatter plot displaying replication timing vs. error, with specific focus on centromeres (blue) and telomeres (orange), alongside the whole-genome data. **(c)** Scatter plot for chromosome 16, zooming in on the common fragile site FRA16D, revealing a continuous error path in mid-to-late replication, near the *WWOX* gene. **(d)** Examples of misfit regions detected by the model across three different chromosomes (3, 6, and 7). Each panel shows the chromosome ideogram, gene locations, and a comparison between the observed data (grey) and model predictions (red), as well the associated error. Notably, large genes overlapping with these misfit regions, including *FHIT* (Chr 3), *PRKN* (Chr 6), and *CNTNAP2* (Chr 7), overlap misfit regions. **(e)** Misfit distribution for common (blue) and rare (pink) fragile sites. The bottom panel shows the misfit fraction relative to site length, while the top panel depicts the length of continuous misfit regions. **(f)** Misfit fraction analysis of the largest genes within fragile sites. **(g)** Scatter plot of replication timing vs. error trajectories for large genes, highlighting error accumulations based on gene size and location within fragile sites. **(h)** Table showing the 10 largest genes misfit by the model across all chromosomes, ranked from largest to smallest (left to right). Genes intersecting fragile sites are highlighted in different colors: blue for common fragile sites, pink for rare fragile sites, and green for genes intersecting both. All plots refer to the H1 cell line, where Repli-seq reads were aligned to hg38.

over, it has been shown that transcription factor-binding sites are key in enhancing DNA replication, as evidenced by studies demonstrating that the presence of these sites within regulatory regions can significantly increase replication efficiency (Turner and Woodworth, 2001). Additionally, previous research has revealed that a significant proportion of replication origins are associated with transcriptional units, particularly at promoter regions, where the density of origins strongly correlates with promoter density (Sequeira-Mendes et al., 2009). This co-evolution of replication and transcription regulatory regions further supports the idea that transcriptional activity not only facilitates replication but also influences the efficiency and organisation of replication origins in mammalian cells. The strong correlation between high origin firing rates and regions of active transcription, open chromatin, and promoters provides further insights into genome-wide coordination of replication and transcription. Notably, putative replication origins are often

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

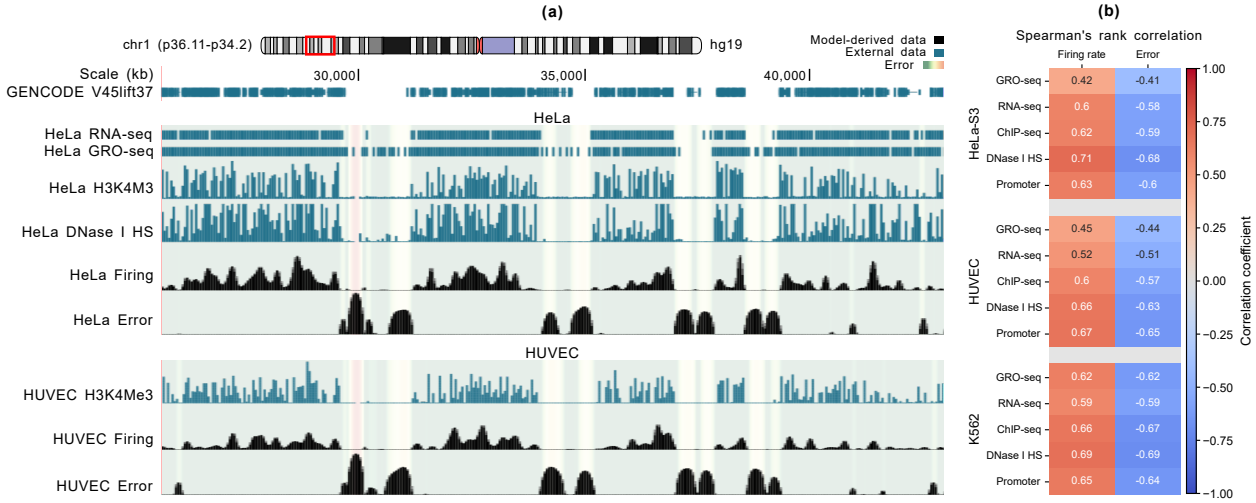


Figure 6. Replication timing discrepancies and firing rate profiles correlate with transcriptional and chromatin data.
(a) Snapshot from the UCSC Genome Browser showing a detailed view of chromosome 1 (p36.11-p34.2) across HUVEC and HeLa lines (hg19). Various tracks compare transcriptional and chromatin data to replication timing misfit magnitude (error) and firing rate profiles obtained from our model (log-scale). The tracks include RNA-seq and GRO-seq (capturing transcription levels and nascent RNA transcription, respectively), as well as ChIP-seq (H3K4Me3 histone modification associated with promoters), DNase I hypersensitivity (indicative of open chromatin regions), and promoter activity. For clarity, the error for each cell line is represented as a translucent heat map across the tracks, with colours ranging from green (indicating a good fit) to yellow/red (poor fit). **(b)** Heatmap displaying the Spearman correlation coefficients between origin firing rates and fit errors with transcriptional and chromatin features for HeLa, HUVEC, and K562 cell lines. All tests returned p-value $< 10^{-15}$.

located within regions of open chromatin, which are enriched in areas that facilitate nucleosome-free regions and early replication (Boulos et al., 2015; Audit et al., 2009), also associated with improved fitting in our model. The presence of promoters in these regions supports this coordinated relationship. In other work, it has been shown that replication initiation often coincides with transcription start sites (Chen et al., 2019). Frequent origin firing in these regions likely ensures early replication initiation in S phase, thereby avoiding potential conflicts with ongoing transcription. This synchronisation between replication and transcription could be a key mechanism to prevent replication stress, particularly during later stages of S phase when transcription is still actively occurring. Furthermore, it has been demonstrated that transcriptionally active euchromatin generally replicates early in S phase, while silent heterochromatin replicates late, with replication timing being both a cause and consequence of chromatin structure (Gilbert, 2002). Under replication stress, this coupling is adjusted, with initiation and termination sites shifting to maintain the balance between replication and transcription, highlighting the intricate coordination that sustains genome integrity.

Our model captures known biological replication patterns, revealing mistfits in transcriptionally active, accessible chromatin regions and heightened misfits around common fragile sites. Beyond aligning with these established relationships, the model also identifies previously uncharted regions where replication timing significantly diverges, providing new insights into areas of potential replication vulnerability. These findings underscore the model's power to both confirm known biology and extend understanding of the intricate relationship between replication timing, chromatin structure, and genome stability.

Discussion

Our modelling efforts have revealed that the intricate dynamics involved in the structural relationship between DNA replication and transcription are deeply encoded within replication timing profiles. By precisely mapping the mathematical link between replication timing and origin firing rates, our model identifies distinctive genomic loci with timing discrepancies, highlighting large genes and fragile sites as regions prone to replication stress. Notably, the model achieves this using Repli-seq data alone, uncovering associations between chromatin openness, active transcription, and timely replication, offering valuable insights into genome integrity.

In our genome-wide simulations, the model effectively captured key replication dynamics, including replication timing, fork directionality, and inter-origin distances. Replication timing was fitted with high precision across most of the genome, with only a few regions where the observations clearly deviate from the simulation. We found that, while misfit distributions varied across different chromosomes and cell lines, late-replicating regions consistently exhibited higher misfit rates. This matches previous findings suggesting these regions are more prone to replication challenges. The analysis also showed a strong negative correlation between firing rates and replication timing misfits, indicating that regions with less frequent origin firing are more susceptible to timing deviations. Additionally, non-coding regions demonstrated a higher frequency of misfits, highlighting potential vulnerabilities in these areas.

In examining fragile sites, we found that the model revealed pronounced replication timing fit errors across extended regions within large genes. Both common and rare fragile sites showed longer continuous chromosomal regions of mistfit, indicating that these misfits are widespread fea-

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tures within these regions and not confined to a specific type of fragile site. A discussion on early replicating fragile sites (ERFSs), which are associated with highly transcribed genes and genome instability, is a topic for future exploration (Barlow et al., 2013).

Importantly, even though our model focuses on kinetic dynamics rather than detailed molecular mechanisms and uses only Repli-seq timing data, we found a strong association between high origin firing rates and regions of active transcription, open chromatin, and promoter-rich areas. This indicates that transcriptionally active regions not only facilitate replication initiation but also help coordinate replication timing. Regions with active transcription and open chromatin exhibited fewer replication timing misfits, likely due to easier access for replication proteins, which outweighs potential conflicts between transcription and replication.

While these findings align with established knowledge and validate the model, they also highlight its potential to reveal additional insights into replication dynamics. The accurate reproduction of known correlations underscores the model's robustness, but the regions of misfit offer the most intriguing avenues for exploration. These misfits, some aligning with known fragile sites and others appearing in novel locations, may refine our understanding of common fragile sites or uncover new regions susceptible to replication stress. This capability transforms the model into a powerful tool for pinpointing interesting genomic sites using Repli-seq data. We hypothesise that these misfits not only provide a finer delineation of fragile sites, but also reveal previously unidentified, smaller regions susceptible to replication stress. Our findings open up exciting possibilities for future research to explore the biological relevance of these misfits and their implications for genome stability and integrity.

One limitation of our modelling approach is the assumption that origin firing rate is independent and reduced to a single parameter, which may not capture the full complexity of origin licensing and activation. However, this simplification allows the model to fit to Repli-seq data rapidly, making it a practical tool for genome-wide analyses. In reality, factors like ORC, Cdc6, and MCM proteins regulate origin licensing, while activation can be influenced by cell cycle checkpoints and stress response pathways (Shechter et al., 2004; Boos and Ferreira, 2019). Integrating additional data sources, such as Hi-C data to examine the three-dimensional structure of chromatin (Gindin et al., 2014), could enhance the model's ability to account for spatial genome organisation and further improve its predictive power.

Secondary DNA structures like R-loops, hairpins, and G-quadruplexes can obstruct replication and cause conflicts with transcription (García-Muse and Aguilera, 2016), impacting replication dynamics in ways our model does not yet capture. Addressing these structures could be a valuable direction for the future expansion of this work. Although the mathematical framework is adaptable to other organisms beyond humans, applying it to species with different genomic architectures, such as *Saccharomyces cerevisiae* with its short genome and precisely located replication origins, will require adjustments to the model's assumptions. In particular, the large genome assumption used in our model may not be directly applicable.

An exciting avenue for expanding this model involves exploring the impact of chemotherapies on DNA replication dynamics, particularly those therapies that target the Replication Stress Response (RSR) pathway and its key signalling proteins (Berti and Vindigni, 2016). By simulating the inhibition of such proteins, the model could provide valuable insights into how these disruptions affect replication timing, origin firing, and potential cell death outcomes (Manic et al., 2018). This expansion could help predict the efficacy of various chemotherapy combinations and offer a cost-effective approach to optimising cancer treatments.

In summary, our model not only validates known relationships between replication dynamics, transcriptional activity, and chromatin state but also uncovers additional regions of replication timing discrepancies that may represent previously unrecognised sites of genomic instability. This dual capability underscores the model's potential as both a predictive tool and a means of discovering novel genomic features. Future investigations focusing on these misfit regions could provide deeper insights into genome regulation and vulnerabilities, enhancing our understanding of replication stress and its role in genomic disorders and cancer.

Data and code availability

The GRO-seq data used in this study were obtained from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under the following accession numbers: GSE62046, GSE94872, and GSE60454.

The source code for the main fitting algorithm, along with the replication timing fit error and origin firing rate bedgraph files, are hosted on the following GitHub repository: https://github.com/fberkemeier/DNA_replication_model.git. The Beacon Calculus simulations were performed using version 1.1.0 of `bcs`, available at https://github.com/MBoemo/bc (Boemo et al., 2020). Additional examples of `bcs` scripts and optimisation algorithms can be found in SN2. For any comments, suggestions, or questions regarding copyright, please contact fp409@cam.ac.uk.

ACKNOWLEDGEMENTS

We thank Prof. Sarah McClelland for her constructive and insightful feedback, which significantly improved the manuscript. We also thank all members of the Boemo lab for their helpful discussions and comments. This work was made possible by the Leverhulme Trust Research Project Grant RPG-2022-028. It was performed using resources provided by the Cambridge Service for Data Driven Discovery (CSD3), operated by the University of Cambridge Research Computing Service (https://www.csd3.cam.ac.uk), and supported by Dell EMC and Intel through Tier-2 funding from the Engineering and Physical Sciences Research Council (capital grant EP/T022159/1) and DiRAC funding from the Science and Technology Facilities Research Council (https://dirac.ac.uk). Additional funding and career support was provided by a Rokos Postdoctoral Associate position at Queens' College Cambridge to Francisco Berkemeier and a fellowship at St John's College Cambridge to Michael A. Boemo.

COMPETING INTERESTS

The authors have declared that no competing interests exist.

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

Anderson, C. J., Talmane, L., Luft, J., Connelly, J., Nicholson, M. D., Verburg, J. C., Pich, O., Campbell, S., Giaisi, M., Wei, P.-C., et al. (2024). Strand-resolved mutagenicity of DNA damage and repair. *Nature*, pages 1–8.

Arbona, J.-M., Kabalane, H., Barbier, J., Goldar, A., Hyrien, O., and Audit, B. (2023). Neural network and kinetic modelling of human genome replication reveal replication origin locations and strengths. *PLOS Computational Biology*, 19(5):e1011138.

Audit, B., Zaghloul, L., Vaillant, C., Chevereau, G., d'Aubenton Carafa, Y., Thermes, C., and Arneodo, A. (2009). Open chromatin encoded in DNA sequence is the signature of 'master'replication origins in human cells. *Nucleic acids research*, 37(18):6064–6075.

Barlow, J. H., Faryabi, R. B., Callén, E., Wong, N., Malhowski, A., Chen, H. T., Gutierrez-Cruz, G., Sun, H.-W., McKinnon, P., Wright, G., et al. (2013). Identification of early replicating fragile sites that contribute to genome instability. *Cell*, 152(3):620–632.

Berti, M. and Vindigni, A. (2016). Replication stress: getting back on track. *Nature structural & molecular biology*, 23(2):103–109.

Blin, M., Le Tallec, B., Nähse, V., Schmidt, M., Brossas, C., Millot, G. A., Prioleau, M.-N., and Debatisse, M. (2019). Transcription-dependent regulation of replication dynamics modulates genome stability. *Nature structural & molecular biology*, 26(1):58–66.

Boemo, M. A., Cardelli, L., and Nieduszynski, C. A. (2020). The beacon calculus: A formal method for the flexible and concise modelling of biological systems. *PLoS computational biology*, 16(3):e1007651.

Boos, D. and Ferreira, P. (2019). Origin firing regulations to control genome replication timing. *Genes*, 10(3):199.

Boulos, R. E., Drillon, G., Argoul, F., Arneodo, A., and Audit, B. (2015). Structural organization of human replication timing domains. *FEBS letters*, 589(20):2944–2957.

Bracci, A. N., Dallmann, A., Ding, Q., Hubisz, M. J., Caballero, M., and Koren, A. (2023). The evolution of the human DNA replication timing program. *Proceedings of the National Academy of Sciences*, 120(10):e2213896120.

Branzei, D. and Foiani, M. (2010). Maintaining genome stability at the replication fork. *Nature reviews Molecular cell biology*, 11(3):208–219.

Briu, L.-M., Maric, C., and Cadoret, J.-C. (2021). Replication stress, genomic instability, and replication timing: A complex relationship. *International Journal of Molecular Sciences*, 22(9):4764.

Chen, Y.-H., Keegan, S., Kahli, M., Tonzi, P., Fenyö, D., Huang, T. T., and Smith, D. J. (2019). Transcription shapes DNA replication initiation and termination in human cells. *Nature structural & molecular biology*, 26(1):67–77.

Cockerill, P. N. (2011). Structure and function of active chromatin and DNase I hypersensitive sites. *The FEBS journal*, 278(13):2182–2210.

Colicino-Murbach, E., Hathaway, C., and Dungrawala, H. (2024). Replication fork stalling in late s-phase elicits nascent strand degradation by DNA mismatch repair. *Nucleic Acids Research*, page gkae721.

Conti, C., Saccà, B., Herrick, J., Lalou, C., Pommier, Y., and Bensimon, A. (2007). Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. *Molecular biology of the cell*, 18(8):3059–3067.

Crawford, G. E., Holt, I. E., Mullikin, J. C., Tai, D., of Health Intramural Sequencing Center†‡, N. I., Green, E. D., Wolfsberg, T. G., and Collins, F. S. (2004). Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proceedings of the National Academy of Sciences*, 101(4):992–997.

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., et al. (2018). The encyclopedia of DNA elements (encode): data portal update. *Nucleic acids research*, 46(D1):D794–D801.

de Moura, A. P., Retkute, R., Hawkins, M., and Nieduszynski, C. A. (2010). Mathematical modelling of whole chromosome replication. *Nucleic acids research*, 38(17):5623–5633.

Donley, N. and Thayer, M. J. (2013). DNA replication timing, genome stability and cancer: late and/or delayed DNA replication timing is associated with increased genomic instability. In *Seminars in cancer biology*, volume 23, pages 80–89. Elsevier.

Durkin, S. G. and Glover, T. W. (2007). Chromosome fragile sites. *Annu. Rev. Genet.*, 41:169–192.

Ehrenhofer-Murray, A. E. (2004). Chromatin dynamics at DNA replication, transcription and repair. *European Journal of Biochemistry*, 271(12):2335–2349.

Franchitto, A. and Pichierri, P. (2014). Replication fork recovery and regulation of common fragile sites stability. *Cellular and molecular life sciences*, 71:4507–4517.

Frankish, A., Carbonell-Sala, S., Diekhans, M., Jungreis, I., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Arnan, C., Barnes, I., et al. (2023). Gencode: reference annotation for the human and mouse genomes in 2023. *Nucleic acids research*, 51(D1):D942–D949.

García-Muse, T. and Aguilera, A. (2016). Transcription–replication conflicts: how they occur and how they are resolved. *Nature reviews Molecular cell biology*, 17(9):553–563.

Gefter, M. L. (1975). DNA replication. *Annual review of biochemistry*, 44(1):45–78.

Gilbert, D. M. (2002). Replication timing and transcriptional control: beyond cause and effect. *Current opinion in cell biology*, 14(3):377–383.

Gindin, Y., Valenzuela, M. S., Aladjem, M. I., Meltzer, P. S., and Bilke, S. (2014). A chromatin structure-based model accurately predicts DNA replication timing in human cells. *Molecular systems biology*, 10(3):722.

Glover, T. W., Berger, C., Coyle, J., and Echo, B. (1984). Dna polymerase α inhibition by aphidicolin induces gaps and breaks at common fragile sites in human chromosomes. *Human genetics*, 67(2):136–142.

Glover, T. W., Wilson, T. E., and Arlt, M. F. (2017). Fragile sites in cancer: more than meets the eye. *Nature Reviews Cancer*, 17(8):489–501.

Guilbaud, G., Rappailles, A., Baker, A., Chen, C.-L., Arneodo, A., Goldar, A., d'Aubenton Carafa, Y., Thermes, C., Audit, B., and Hyrien, O. (2011). Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. *PLoS computational biology*, 7(12):e1002322.

Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., Weaver, M., Dorschner, M. O., Gartler, S. M., and Stamatoyannopoulos, J. A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*, 107(1):139–144.

Herrick, J., Jun, S., Bechhoefer, J., and Bensimon, A. (2002). Kinetic model of DNA replication in eukaryotic organisms. *Journal of molecular biology*, 320(4):741–750.

Huang, X., Gao, X., Li, W., Jiang, S., Li, R., Hong, H., Zhao, C., Zhou, P., Chen, H., Bo, X., et al. (2019). Stable H3K4me3 is associated with transcription initiation during early embryo development. *Bioinformatics*, 35(20):3931–3936.

Hulke, M. L., Massey, D. J., and Koren, A. (2020). Genomic methods for measuring DNA replication dynamics. *Chromosome Research*, 28(1):49–67.

Jun, S. and Bechhoefer, J. (2005). Nucleation and growth in one dimension. II. application to DNA replication kinetics. *Physical Review E*, 71(1):011909.

Jun, S. and Rhind, N. (2008). Just-in-time DNA replication. *Physics*, 1.

Jun, S., Zhang, H., and Bechhoefer, J. (2005). Nucleation and growth in one dimension. I. the generalized kolmogorov-johnson-mehl-avrami model. *Physical Review E*, 71(1):011908.

Kadonaga, J. T. (1998). Eukaryotic transcription: an interlaced network of transcription factors and chromatin-modifying machines. *Cell*, 92(3):307–313.

Kaushal, S. and Freudenreich, C. H. (2019). The role of fork stalling and DNA structures in causing chromosome fragility. *Genes, Chromosomes and Cancer*, 58(5):270–283.

Knott, S. R., Viggiani, C. J., and Aparicio, O. M. (2009). To promote and protect: coordinating DNA replication and transcription for genome stability. *Epigenetics*, 4(6):362–365.

Kogoma, T. (1997). Stable DNA replication: interplay between DNA replication, homologous recombination, and transcription. *Microbiology and Molecular Biology Reviews*, 61(2):212–238.

Kolmogorov, A. (1937). On the static theory of crystallization in metals. *Bull Acad Sci USSR, Phys Ser*, 1:335.

Koren, A., Handsaker, R. E., Kamitaki, N., Karlić, R., Ghosh, S., Polak, P., Eggan, K., and McCarroll, S. A. (2014). Genetic variation in human DNA replication timing. *Cell*, 159(5):1015–1026.

Kumar, R., Nagpal, G., Kumar, V., Usmani, S. S., Agrawal, P., and Raghava, G. P. (2019). Humcfs: a database of fragile sites in human chromosomes. *BMC genomics*, 19:1–8.

Laird, C., Jaffe, E., Karpen, G., Lamb, M., and Nelson, R. (1987). Fragile sites in human chromosomes as regions of late-replicating DNA. *Trends in Genetics*, 3:274–281.

Le Beau, M. M., Rassool, F. V., Neilly, M. E., Espinosa III, R., Glover, T. W., Smith, D. I., and McKeithan, T. W. (1998). Replication of a common fragile site, fra3b, occurs late in s phase and is delayed further upon induction: implications for the mechanism of fragile site induction. *Human molecular genetics*, 7(4):755–761.

Leonard, A. C. and Méchali, M. (2013). DNA replication origins. *Cold Spring Harbor perspectives in biology*, 5(10):a010116.

Li, S. and Wu, X. (2020). Common fragile sites: protection and repair. *Cell & bioscience*, 10:1–9.

Lopes, R., Agami, R., and Korkmaz, G. (2017). GRO-seq, a tool for identification of transcripts regulating gene expression. *Promoter Associated RNA: Methods and Protocols*, pages 45–55.

Lu, R. and Pickett, H. A. (2022). Telomeric replication stress: the beginning and the end for alternative lengthening of telomeres cancers. *Open Biology*, 12(3):220011.

Manic, G., Sistigu, A., Corradi, F., Musella, M., De Maria, R., and Vitale, I. (2018). Replication stress response in cancer stem cells as a target for chemotherapy. In *Seminars in cancer biology*, volume 53, pages 31–41. Elsevier.

Marchal, C., Sima, J., and Gilbert, D. M. (2019). Control of DNA replication timing in the 3D genome. *Nature Reviews Molecular Cell Biology*, 20(12):721–737.

Marguerat, S. and Bähler, J. (2010). RNA-seq: from technology to biology. *Cellular and molecular life sciences*, 67:569–579.

Maric, C. and Prioleau, M.-N. (2010). Interplay between DNA replication and gene expression: a harmonious coexistence. *Current opinion in cell biology*, 22(3):277–283.

Massey, D. J. and Koren, A. (2022). Telomere-to-telomere human DNA replication timing profiles. *Scientific Reports*, 12(1):9560.

Mirkin, E. V. and Mirkin, S. M. (2007). Replication fork stalling at natural impediments. *Microbiology and molecular biology reviews*, 71(1):13–35.

Müller, C. A. and Nieduszynski, C. A. (2012). Conservation of replication timing reveals global and local regulation of replication origin activity. *Genome research*, 22(10):1953–1962.

Müller, C. A. and Nieduszynski, C. A. (2017). DNA replication timing influences gene expression level. *Journal of Cell Biology*, 216(7):1907–1914.

Palakodeti, A., Han, Y., Jiang, Y., and Le Beau, M. M. (2004). The role of late/slow replication of the fra16d in common fragile site induction. *Genes, Chromosomes and Cancer*, 39(1):71–76.

Pepke, S., Wold, B., and Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nature methods*, 6(Suppl 11):S22–S32.

Quinet, A., Carvajal-Maldonado, D., Lemacon, D., and Vindigni, A. (2017). DNA fiber analysis: mind the gap! *Methods in enzymology*, 591:55–82.

Retkute, R., Nieduszynski, C. A., and De Moura, A. (2012). Mathematical modeling of genome replication. *Physical Review E*, 86(3):031916.

Rhind, N. and Gilbert, D. M. (2013). DNA replication timing. *Cold Spring Harbor perspectives in biology*, 5(8):a010132.

Schwartz, M., Zlotorynski, E., and Kerem, B. (2006). The molecular basis of common and rare fragile sites. *Cancer letters*, 232(1):13–26.

Sequeira-Mendes, J., Díaz-Uriarte, R., Apedaile, A., Huntley, D., Brockdorff, N., and Gómez, M. (2009). Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS genetics*, 5(4):e1000446.

Shechter, D., Costanzo, V., and Gautier, J. (2004). Atr and atm regulate the timing of DNA replication origin firing. *Nature cell biology*, 6(7):648–655.

Sinai, M. I.-T. and Kerem, B. (2018). DNA replication stress drives fragile site instability. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 808:56–61.

Smith, D. I., Zhu, Y., McAvoy, S., and Kuhn, R. (2006). Common fragile sites, extremely large genes, neural development and cancer. *Cancer letters*, 232(1):48–57.

Tabancay Jr, A. P. and Forsburg, S. L. (2006). Eukaryotic DNA replication in a chromatin context. *Current topics in developmental biology*, 76:129–184.

Técher, H., Koundrioukoff, S., Azar, D., Wilhelm, T., Carignon, S., Brison, O., Debatisse, M., and Le Tallec, B. (2013). Replication dynamics: biases and robustness of DNA fiber analysis. *Journal of molecular biology*, 425(23):4845–4855.

Temme, N. M. (2010). Error functions, dawson's and fresnel integrals.

Thurman, R. E., Day, N., Noble, W. S., and Stamatoyannopoulos, J. A. (2007). Identification of higher-order functional domains in the human encode regions. *Genome research*, 17(6):917–927.

Turner, W. J. and Woodworth, M. E. (2001). DNA replication efficiency depends on transcription factor-binding sites. *Journal of Virology*, 75(12):5638–5645.

Tyurin, A. N. (2002). Quantization, classical and quantum field theory and theta-functions. *arXiv*

bioRxiv preprint doi: https://doi.org/10.1101/2024.11.25.625090; this version posted November 25, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

preprint math/0210466.

Waga, S. and Stillman, B. (1998). The DNA replication fork in eukaryotic cells. *Annual review of biochemistry*, 67(1):721–751.

Wang, W., Klein, K. N., Proesmans, K., Yang, H., Marchal, C., Zhu, X., Borrman, T., Hastie, A., Weng, Z., Bechhoefer, J., et al. (2021). Genome-wide mapping of human DNA replication by optical replication mapping supports a stochastic model of eukaryotic replication. *Molecular cell*, 81(14):2975–2988.

Woo, Y. H. and Li, W.-H. (2012). DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nature communications*, 3(1):1004.

Young, M. D., Willson, T. A., Wakefield, M. J., Trounson, E., Hilton, D. J., Blewitt, M. E., Oshlack, A., and Majewski, I. J. (2011). ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic acids research*, 39(17):7415–7427.

Zhao, P. A., Sasaki, T., and Gilbert, D. M. (2020). High-resolution Repli-seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome biology*, 21:1–20.

Supplementary Information

Supplementary Note 1: Mathematical notes

1.1. Expected time of replication

Without loss of generality, we assume a ring network (periodic DNA) to enforce symmetry of replication with respect to a focal origin. In a large genome, this periodic assumption has minimal influence across most regions, apart from the chromosome ends.

Let T be the time a site takes to either fire (if it is a replication origin) or be replicated by an incoming fork. We can think of T as an explicit function of the origin firing times A_i, where $A_i \overset{\text{iid}}{\sim} \text{Exp}(f)$. In particular, $\mathbb{E}[A_i] = 1/f$. We index each site by its distance from the origin of interest, given by $|i|$. Notice that $i = 0$ corresponds to the focal origin, and v is interpreted as the number of replicated sites per time unit. We have

$$T = \min_i \{A_i + |i|/v\} \tag{S1}$$

since it takes time $|i|/v$ for a replication fork initiated at site i to reach the origin of interest. Then,

$$P(T > t) = \prod_i P(A_i > t - |i|/v) = \prod_i \min\{1, \exp(-f(t - |i|/v))\} \tag{S2}$$

since $A_i > 0$ and $A_i \overset{\text{iid}}{\sim} \text{Exp}(f)$. Hence, the expectation of the replication time for any one site is given by

$$\mathbb{E}[T] = \int_0^\infty P(T > t)\, dt = \int_0^\infty \prod_i \min\{1, \exp(-f(t - |i|/v))\}\, dt. \tag{S3}$$

This integral can be partitioned across each interval for which $|i| \le vt \le |i+1|$. Within these intervals, the integrands adopt the form ae^{-bt}, thereby permitting analytical evaluation. A few particular cases include:

- One origin ($n = 1$):

$$\mathbb{E}[T; 1] = \int_0^\infty e^{-ft} dt = \frac{1}{f} \tag{S4}$$

- Two origins ($n = 2$):

$$\mathbb{E}[T; 2] = \int_0^{\frac{1}{v}} e^{-ft} dt + \int_{\frac{1}{v}}^\infty e^{-f(2t-1/v)} dt = \frac{1}{f}\left(1 - \frac{1}{2}e^{-\frac{f}{v}}\right) \tag{S5}$$

- Three origins ($n = 3$):

$$\mathbb{E}[T; 3] = \int_0^{\frac{1}{v}} e^{-ft} dt + \int_{\frac{1}{v}}^\infty e^{-f(3t-2/v)} dt = \frac{1}{f}\left(1 - \frac{2}{3}e^{-\frac{f}{v}}\right) \tag{S6}$$

- Four origins ($n = 4$):

$$\mathbb{E}[T; 4] = \int_0^{\frac{1}{v}} e^{-ft} dt + \int_{\frac{1}{v}}^{\frac{2}{v}} e^{-f(3t-2/v)} dt + \int_{\frac{2}{v}}^\infty e^{-f(4t-4/v)} dt = \frac{1}{f}\left(1 - \frac{1}{12}e^{-4\frac{f}{v}} - \frac{2}{3}e^{-\frac{f}{v}}\right) \tag{S7}$$

where $\mathbb{E}[T; n] \equiv \mathbb{E}[T]$ for each n. In the general case, the result depends on the parity of n. When n is odd, for each k, there are 2 origins at a distance of $k = 1, 2, \ldots, (n-1)/2$ from the origin of interest. Adding up these distances leads to

$$\mathbb{E}[T; n_{\text{odd}}] = \sum_{k=0}^{(n-3)/2} \int_{k/v}^{(k+1)/v} e^{-f((2k+1)t - k(k+1)/v)}\, dt + \int_{(n-1)/(2v)}^\infty e^{-f(nt - (n-1)(n+1)/(4v))}\, dt, \tag{S8}$$

where the last term is just the $k = (n-1)/2$ term of the sum with the upper limit replaced by ∞. Solving the integrals yields

$$\mathbb{E}[T; n_{\text{odd}}] = \frac{1}{f}\left[\sum_{k=0}^{(n-3)/2} \frac{e^{-fk^2/v} - e^{-f(k+1)^2/v}}{2k+1} + \frac{e^{-f(n-1)^2/(4v)}}{n}\right]. \tag{S9}$$

When n is even, for each k there are 2 origins at a distance of $k = 1, 2, \ldots, (n-2)/2$, and then there is 1 origin at a distance of $n/2$. Again, we add up the distances, each twice, but since there is only one origin at a distance of $n/2$, the very last distance sum is $n^2/4$. So, we get

$$\mathbb{E}[T; n_{\text{even}}] = \sum_{k=0}^{(n-2)/2} \int_{k/v}^{(k+1)/v} e^{-f((2k+1)t - k(k+1)/v)} dt + \int_{n/(2v)}^{\infty} e^{-f(nt - n^2/(4v))} dt. \tag{S10}$$

Solving the integrals yields

$$\mathbb{E}[T; n_{\text{even}}] = \frac{1}{f} \left[\sum_{k=0}^{(n-2)/2} \frac{e^{-fk^2/v} - e^{-f(k+1)^2/v}}{2k+1} + \frac{e^{(-fn^2/(4v))}}{n} \right]. \tag{S11}$$

Using the ceiling function $\lceil \cdot \rceil$ to handle parity, a general expression for each origin, and any n, is

$$\mathbb{E}[T; n] \equiv \frac{1}{f} \left[\sum_{k=0}^{\lceil (n-3)/2 \rceil} \frac{e^{-fk^2/v} - e^{-f(k+1)^2/v}}{2k+1} + \frac{e^{-f(\lceil (n-1)/2 \rceil)^2/v}}{n} \right]. \tag{S12}$$

In particular,

$$\mathbb{E}[T; \infty] \equiv \lim_{n \to \infty} \mathbb{E}[T; n] = \frac{1}{f} \sum_{k=0}^{\infty} \frac{e^{-fk^2/v} - e^{-f(k+1)^2/v}}{2k+1} \tag{S13}$$

which is Eq. (4). Eq. (8) arises from a similar reasoning, achieved by expressing the product of exponentials as a single exponential of sums. Although the series $\mathbb{E}[T; n]$ converges for $f > 0$, its closed-form expression is not known. If we rescale time $\tilde{T} \equiv fT$, $\tilde{t} \equiv ft$, and define $x \equiv f/v$, we may rewrite Eq. (S12) in a more compact, non-dimensional form

$$\mathbb{E}[\tilde{T}; n] \equiv \sum_{k=0}^{\lceil (n-3)/2 \rceil} \frac{e^{-xk^2} - e^{-x(k+1)^2}}{2k+1} + \frac{e^{-x(\lceil (n-1)/2 \rceil)^2}}{n}. \tag{S14}$$

As $n \to \infty$, we have

$$\mathbb{E}[\tilde{T}; \infty] \equiv \lim_{n \to \infty} \mathbb{E}[\tilde{T}; n] = \sum_{k=0}^{\infty} \frac{e^{-xk^2} - e^{-x(k+1)^2}}{2k+1} = \sum_{k \in \mathbb{Z}} \frac{e^{-xk^2}}{1 - 4k^2}. \tag{S15}$$

A few interesting observations can be made regarding the upper bounds of this limit.

1.2. On Dawson function estimates

The series $g(x) \equiv \mathbb{E}[\tilde{T}; \infty]$ is related to the family of theta functions (Tyurin, 2002), allowing us to express it in terms of

$$\vartheta(x) = \sum_{k \in \mathbb{Z}} e^{-\pi (xk)^2} \tag{S16}$$

which satisfies $\vartheta(1/x) = x\vartheta(x)$. From Eq. (S15), g satisfies

$$g(x) + 4g'(x) = \sum_{k \in \mathbb{Z}} e^{-xk^2} = \vartheta(\sqrt{x/\pi}), \tag{S17}$$

and thus

$$g(x) = e^{-x/4} \int_0^{x/4} e^y \, \vartheta(2\sqrt{y/\pi}) \, dy. \tag{S18}$$

In particular, for small x we have

$$g(x) = \sqrt{\pi} D_+(\sqrt{x}/2) + O(xe^{-\pi^2/x}) \tag{S19}$$

where

$$D_+(z) = e^{-z^2} \int_0^z e^{t^2} \, dt = \frac{1}{2} \sum_{n=0}^{\infty} \frac{(-1)^n n!}{(2n+1)!} (2z)^{2n+1} \tag{S20}$$

is the Dawson function (Temme, 2010). A less accurate estimate is then $g(x) = \sqrt{\pi x}/2 + O(x^{3/2})$. Various upper bounds may also be obtained this way. Reverting the change of variables, we get $\mathbb{E}[T; \infty] \simeq \frac{1}{2}\sqrt{\frac{\pi}{fv}}$, as in Eq. (5).

Supplementary Note 2: Computational methods and data

2.1. Beacon Calculus model

As discussed in Boemo et al. (2020), a simplistic model of DNA replication using `bcs` consists of three core process definitions: replication origins (`ORI`), left-moving forks (`FL`), and right-moving forks (`FR`). The origins are positioned along the chromosome of length `L`. Each of these three processes possesses a unique parameter, denoted as `i`, which is assumed to be a specific position on the chromosome between 1 and `L`. In addition to this, the origins have one more parameter: a replication initiation rate, known as `fire`, or f in our model (generalised models may also include licensing probabilities). To monitor which positions on the chromosome have already undergone replication, the model uses markers called beacons. Upon the replication of position `i` by a fork, a beacon is dispatched on the `chr` channel with parameter `i`.

The following is an example of the `bcs` script with 10 replication origins equally spaced over 100 sites

```
// DNA Replication

// Variables
// Chromosome length
L = 100;
// Fast rate
fast = 100000;
// Fork velocity
v = 1.4;

// Process definitions
ORI[i,fire] = {~chr?[i],fire}.(FL[i]||FR[i]);

FR[i] = {chr![i],fast}.[i < L] -> {~chr?[i+1],v}.FR[i+1];
FL[i] = {chr![i],fast}.[i > 0] -> {~chr?[i-1],v}.FL[i-1];

// Process initiation
ORI[1,0.06048832790213383] || ORI[12,0.002045183033099289]
 || ORI[23,0.0012753405213046796] || ORI[34,0.0011945930278953077]
 || ORI[45,0.001035526093646997] || ORI[56,0.0011165358858784408]
 || ORI[67,0.002560893635329413] || ORI[78,0.003411336829553979]
 || ORI[89,0.0022730688407988954] || ORI[100,0.0038028859830789045];

// End
```

A periodic version of DNA replication can be achieved by changing both `FR` and `FL` process definitions to

```
FR[i] = {chr![i],fast}.(([i<L] -> {~chr?[i+1],v}.FR[i+1]) || ([i==L] -> {~chr?[0],v}.FR[0]));
FL[i] = {chr![i],fast}.(([i>0] -> {~chr?[i-1],v}.FL[i-1]) || ([i==0] -> {~chr?[L],v}.FL[L]));
```

2.2. Fitting algorithm

The following code presents the main fitting function, `fitfunction`, used in the fitting algorithm described in previous sections. It provides an efficient way of computing Eq. (8) to mimic `bcs` simulations for non-uniform firing rates. `fitfunction` accepts four arguments: `list` (a data vector with the RT profile of the entire genome), `v0` (average fork speed, usually set to 1.4 kb/min), and `st0` (parameter R, as discussed before). The first guess `x00` is then constructed based on `list`, by Eq. (6). We use an adapted version of `np.roll()`. Data was processed via the Python extension pyBigWig (Ryan et al., 2021). See https://github.com/fberkemeier/DNA_replication_model.git for further details.

```
# Import dependencies
import cProfile
import math
from time import monotonic
from typing import Any
import numpy as np

# Main function
def fitfunction(list, v0, st0):

        timel = list
        v = v0
        st = st0
        exp_v = np.exp(-1/v)
        x00 = np.array([(math.pi/(4*v))*i**(-2) for i in timel])

        # VECTORIZED APPROACH

        def  fast_roll_add(dst, src, shift):
            dst[shift:] += src[:-shift]
            dst[:shift] += src[-shift:]

        def  fp(x, L, v):
            n = len(x)
            y = np.zeros(n)
            last_exp_2_raw = np.zeros(n)
            last_exp_2 = np.ones(n)
            unitary = x.copy()
            for k in range(L+1):
                if k != 0:
                    fast_roll_add(unitary, x, k)
                    fast_roll_add(unitary, x, -k)
                exp_1_raw = last_exp_2_raw
                exp_1 = last_exp_2
                exp_2_raw = exp_1_raw + unitary / v
                exp_2 = np.exp(-exp_2_raw)

                # Compute the weighted sum for each j and add to the total
                y += (exp_1 - exp_2) / unitary

                last_exp_2_raw = exp_2_raw
                last_exp_2 = exp_2
            return y

        def  fitf(time, lst, x0, j):
                return x0[j] * (lst[j]/time[j])**(2)

        def  cfit(time, lst, x0):
            result = np.empty_like(x0)
            for j in range(len(x0)):
                if fitf(time, lst, x0, j) < 10**(-20):
                    result[j] = 10**(-20)
                elif abs(time[j] - lst[j]) < .5:
                    result[j] = x0[j]
                else:
                    result[j] = fitf(time, lst, x0, j)
            return result

        xs = x00
        my_list = [':.20f'.format(i) for i in xs]

        return my_list
```

2.3. Data mappability

Repli-seq data often face mappability issues, particularly in regions with repetitive sequences or low complexity, where short DNA reads cannot be accurately mapped (Hansen et al., 2010; Zhao et al., 2020).

Based on data from Hansen et al. (2010), these regions of low or problematic mappability account for approximately 20% of the whole genome and around 25% of high-error regions (defined as those with errors exceeding 10^2 min), highlighting their relevance in areas prone to replication timing errors. The mean size of these gaps is approximately 42.37 kb (Figure S1). On average, we observed a phi coefficient of 0.21 when comparing high-error regions and problematic loci, indicating a weak positive correlation between the two. This coefficient, derived from a contingency table, suggests that while there is some overlap between high-error and masked regions, the correlation is not strong. Despite this overlap, mappability issues do not significantly affect overall replication timing analyses, as the majority of high-error regions occur in well-mapped genomic areas, ensuring the reliability of the data.

Given the low phi coefficient, we do not exclude these data from our analysis, since the presence of low mappability regions does not appear to be a major factor influencing replication timing errors, allowing us to retain these data in our analysis without compromising its validity.

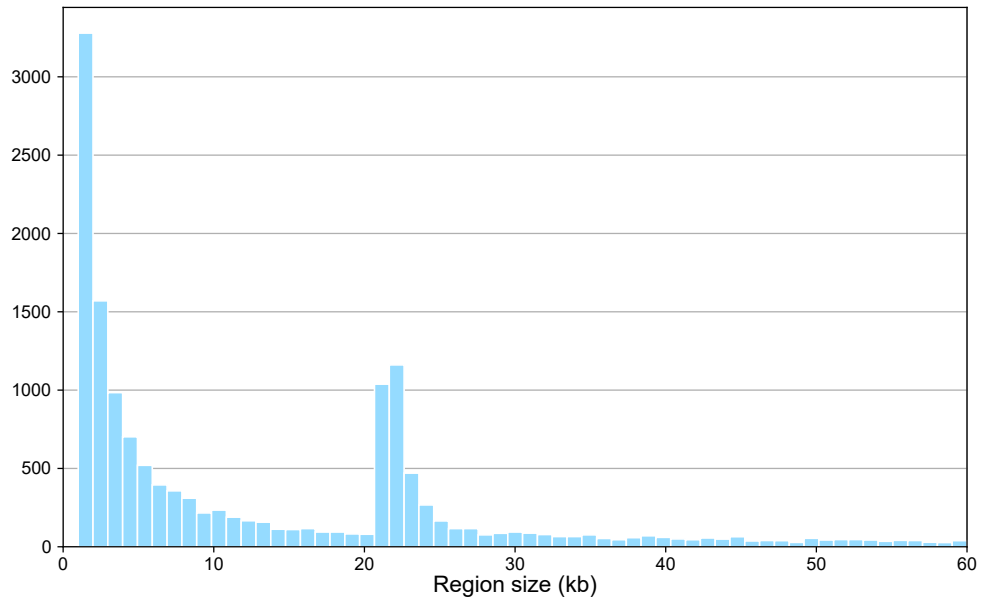


Figure S1. Distribution of problematic mappability region sizes.
Histogram showing the distribution of region sizes with low or problematic mappability (in kilobases) across the genome. These regions are excluded from replication timing analyses due to difficulties in accurately mapping sequencing reads. The majority of these regions are small, with peaks around 1-5 kb and another noticeable peak around 20 kb. The mean size of these regions is approximately 42.37 kb.

2.4. Data correlations

Here, we present a comparison of different statistical tests applied to the datasets discussed in the main text. This analysis evaluates the relationships between replication timing error, firing rates, and transcriptional or chromatin features, providing insights into the suitability and results of Pearson, Spearman rank, and Kendall's tau tests for these data.

Pearson, Spearman rank, and Kendall's tau offer distinct advantages based on the nature of the data and relationships analyzed. Pearson is suited for continuous, normally distributed data with linear relationships, while Spearman rank excels with non-linear or ordinal data by capturing monotonic trends through ranked values. Kendall's tau is particularly effective for smaller datasets, using concordant and discordant pairs to measure associations. Given the non-linear and ranked nature of replication metrics, Spearman rank is ideal for our analysis. Figure S2 shows the correlations between replication timing error, firing rates, and transcriptional or chromatin features, demonstrating the relevance of these tests to our data.

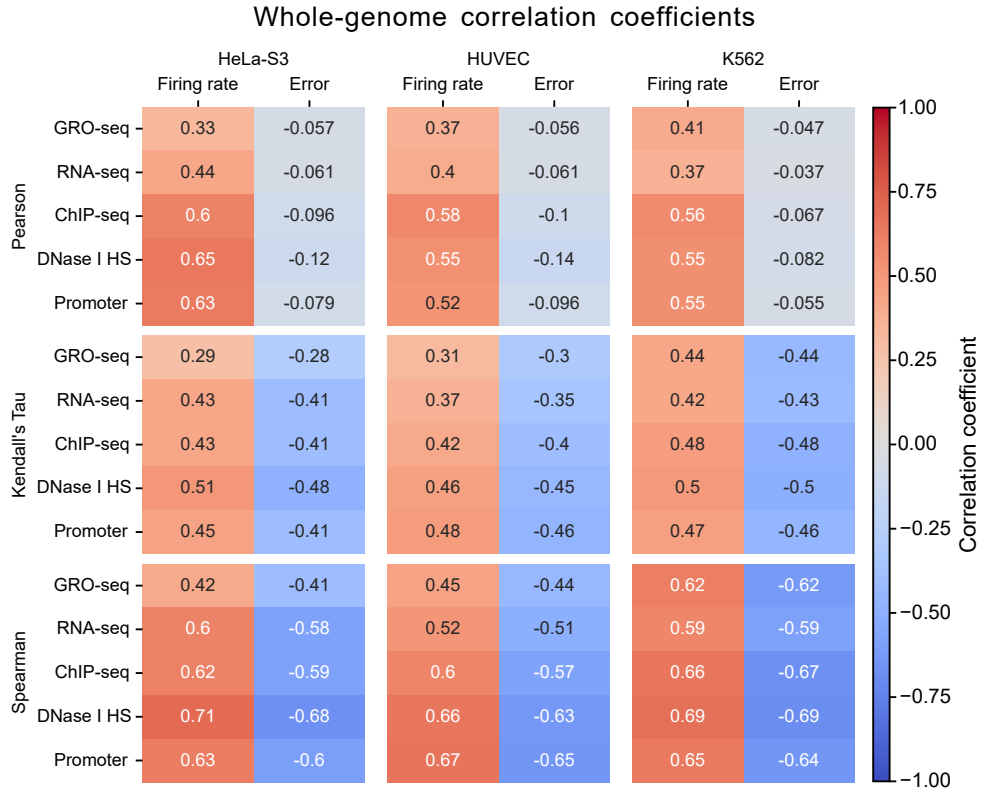


Figure S2. Correlations between replication, transcription and chromatin data.
Heatmap displaying the Spearman, Kendall's Tau, and Pearson correlation coefficients between origin firing rates and fit errors with transcriptional and chromatin features for HeLa, HUVEC, and K562 cell lines. All tests returned p-value $< 10^{-15}$.

References

Boemo, M. A., Cardelli, L., and Nieduszynski, C. A. (2020). The beacon calculus: A formal method for the flexible and concise modelling of biological systems. *PLoS computational biology*, 16(3):e1007651.

Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., Weaver, M., Dorschner, M. O., Gartler, S. M., and Stamatoyannopoulos, J. A. (2010). Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*, 107(1):139–144.

Ryan, D., Eraslan, G., Grüning, B., Silva, R., Marks, P., and Ramirez, F. (2021). pybigwig. deeptools/pybigwig: 0.3.17 (version 0.3.5).

Temme, N. M. (2010). Error functions, dawson's and fresnel integrals.

Tyurin, A. N. (2002). Quantization, classical and quantum field theory and theta-functions. *arXiv preprint math/0210466*.

Zhao, P. A., Sasaki, T., and Gilbert, D. M. (2020). High-resolution Repli-seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome biology*, 21:1–20.