nature communications

Article

https://doi.org/10.1038/s41467-025-59991-w

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

Received: 20 December 2024

Accepted: 12 May 2025

Published online: 19 May 2025

Check for updates

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

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 infers firing rate distributions from Repli-seq timing data across multiple cell lines, enabling a genome-wide comparison between predicted and observed replication. Notably, regions where the model and data diverge often overlap fragile sites and long genes, highlighting the influence of genomic architecture on replication dynamics. Conversely, regions of strong concordance are associated with open chromatin and active promoters, where elevated firing rates facilitate timely fork progression and reduce replication stress. In this work, we provide a valuable framework for exploring the structural interplay between replication timing, transcription, and chromatin organisation, offering insights into the mechanisms underlying replication stress and its implications for genome stability and disease.

Accurate DNA replication is essential for faithfully duplicating genetic information, ensuring its preservation for future generations[1]. In humans, replication occurs during S phase when multiple discrete chromosomal sites, termed origins of replication[2], fire to initiate bidirectional replication forks—molecular machines that traverse the chromosome and replicate DNA[3]. These forks move in opposite directions, progressing until they encounter another fork, reach a chromosome end (Fig. 1a), or face an obstacle (e.g., a bound protein or transcription complex[4]). 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[5].

Replication timing refers to the time at which a specific locus either fires (if an origin) or is passively replicated by an incoming fork. These timing profiles are closely associated with various chromatin structures[6], as well as gene expression[7] and replication stresses[8]. Furthermore, timing is linked to genetic variation[9] and cancer (where late or delayed replication often correlates with increased genomic instability[10]). 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[11]. 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[12].

Replication, transcription, and chromatin organisation are also intricately inter-connected, with each influencing the other[13–15]. In particular, chromatin remodelling regulates the accessibility of regulatory factors, influencing both gene expression and replication. Open chromatin is strongly linked to transcriptional activity and plays a crucial role in replication timing[16,17]. 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[18], ChIP-seq[19], GRO-seq[20], and a database of fragile sites (HumCFS[21]), 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. Our model acts as a

[1]Department of Pathology, University of Cambridge, Cambridge, UK. [2]Department of Genetics, University of Cambridge, Cambridge, UK. [3]Sir William Dunn School of Pathology, University of Oxford, Oxford, UK. ✉e-mail: fp409@cam.ac.uk; mb915@cam.ac.uk

https://doi.org/10.1038/s41467-025-59991-w

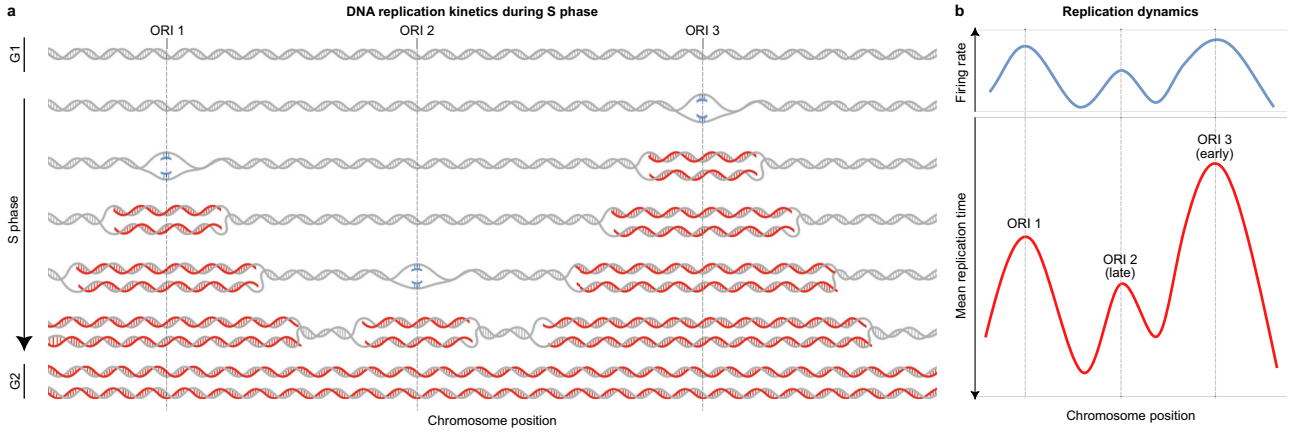



**Fig. 1 | A kinetic model of DNA replication. a** Replication initiates at specific origins that are licensed by the end of G1 phase. During S phase, replication forks progress bidirectionally from origins, passively replicating DNA until they merge with forks from adjacent origins or reach chromosome ends to complete replication and enter G2. In this example, three origins (ORIs 1, 2, and 3) fire at different times, with nascent DNA strands shown in red. At the end of replication, two identical copies of the original template are formed. **b** Illustration of 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 cell population, with peaks corresponding to potential origins. ORI 2 is 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.[70].

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 experience 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[22–25], 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[26], 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[27], we model replication forks and origins as a concurrent system, simulating their behaviour across the genome. In this work, we demonstrate how a theoretical description of replication timing uncovers key links between timing, genomic stability, and other essential genomic processes.

## Results

### An equation for DNA replication timing

We begin by introducing our stochastic framework for replication timing, which is fully detailed in the Methods and in Supplementary Note 1. We aim to identify and quantify genomic regions where replication timing deviates from theoretical 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 (Fig. 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 our framework, the genome is divided into 1 kb segments (sites). Each site $j$ can fire as an origin at a rate $f_j$, while replication forks progress at a constant speed $v$. Concretely, the waiting time for each site's origin to fire follows an exponential distribution with parameter $f_j$. Let $T_j$ be the time at which site $j$ is replicated, either by firing as an origin itself or by being passively replicated by an incoming fork. The expected replication time at any site $j$ is then obtained by weighting the contributions from all potential origins, leading to the following closed-form expression

$$\mathbb{E}[T_j] = \sum_{k=0}^{R} \frac{e^{-\sum_{|i|\le k}(k-|i|)f_{j+i}/v} - e^{-\sum_{|i|\le k}(k+1-|i|)f_{j+i}/v}}{\sum_{|i|\le k} f_{j+i}} \tag{1}$$

where the indices $\{j \pm i\}$ cover neighbouring loci within a chosen radius of influence $R$, i.e., the distance within which neighbouring origins are assumed to affect the timing of a focal origin. Equation (1) enables us to infer the stochastic model's firing rates $\{f_j\}$ from timing data (e.g., Repli-seq), generating a best-fit timing profile for the entire genome that can then be compared with the observed measurements. Regions exhibiting significant discrepancies (misfits) can indicate replication stress or other biological factors not captured by the model. In the following sections, we apply this model to different human cell lines, demonstrating how it reproduces global replication patterns and highlights specific genomic loci that may warrant deeper investigation.

### 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. (1). Then, replication is simulated using Beacon Calculus (`bcs`), a concise process algebra ideal for concurrent systems (see Supplementary Note 2). Finally, we explore patterns of replication seen after averaging 500 simulations for each of the 11 lines (Fig. 2a).

https://doi.org/10.1038/s41467-025-59991-w

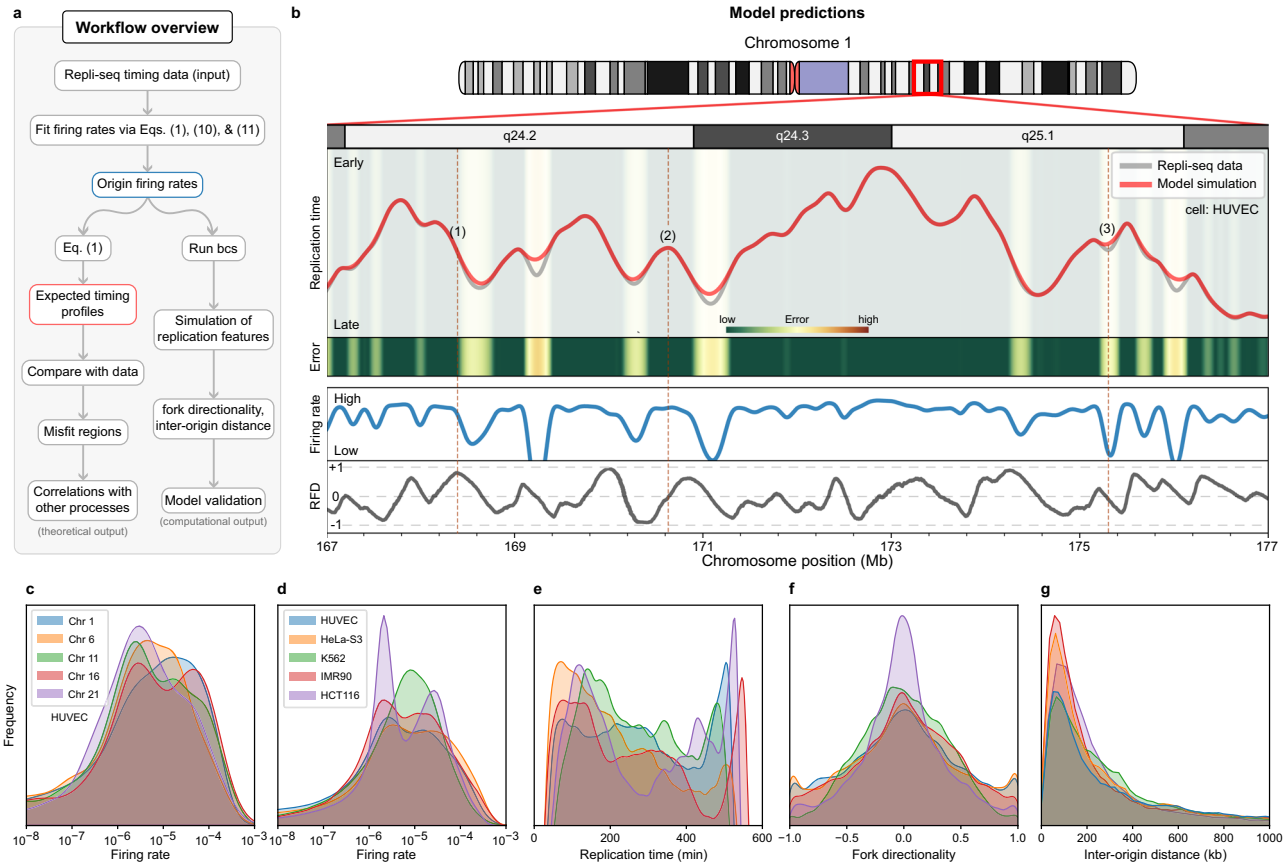



**Fig. 2 | Predicting genome-wide features of replication. a** Overview of the main model and analysis. Starting with Repli-seq timing data, origin firing rates are fitted through Eqs. (1), (10), and (11). These rates generate expected timing profiles for comparison with experimental data to identify 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. Here we see the replication timing of both experimental and simulated data, and the magnitude of the misfit (error) for replication timing in a region where replication forks often stall; this leads to elevated errors that the model struggles to capture accurately. We also show the 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 -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 values are omitted to emphasize relative shapes and distributions rather than absolute magnitudes.

We begin by comparing experimental timing profiles to those obtained from Eq. (1). 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 Fig. 2b. As expected, some regions replicate early (e.g., around 173 Mb) and others late (e.g., around 171 Mb). Overall, the model's predictions agree well with the experimental data (see Methods).

We focus on the regions with high misfit error (shaded in yellow and red in Fig. 2b), where the assumption of constant fork speed in Eq. (1) leads the model to predict earlier replication times than are observed experimentally. Because a constant speed imposes an inherent upper bound on how quickly replication can transition from early to late (see Supplementary Note 2.2.2), any steeper or delayed transitions—potentially arising from fork stalling or local inefficiencies—remain unmatched by the model. These high-misfit zones thus highlight loci where forks appear to slow or stall beyond our simplified assumptions, flagging potential replication stress hotspots for more detailed study.

While firing rates are directly inferred from Eq. (1), 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 (Fig. 2b).

To validate the model, we examine global distributions of multiple features. Despite little variation in firing in HUVECs (Fig. 2c), HCT116 exhibits a pronounced bimodal pattern, likely driven by differences in data sources[28] (Fig. 2d, e), 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 (Fig. 2f). This pattern aligns with recent quantifications of fork directionality in human cells[29]. While determining inter-origin distances (IODs) is straightforward from our simulations, doing so from DNA-fibre experiments remains challenging due to technical limitations and potential biases[30]. Nevertheless, simulations show a concentration of IODs within the commonly observed range of 100–200 kb[31] (Fig. 2g).

Although these results validate the model against established metrics, its broader ability to simulate other features, like replicon

https://doi.org/10.1038/s41467-025-59991-w

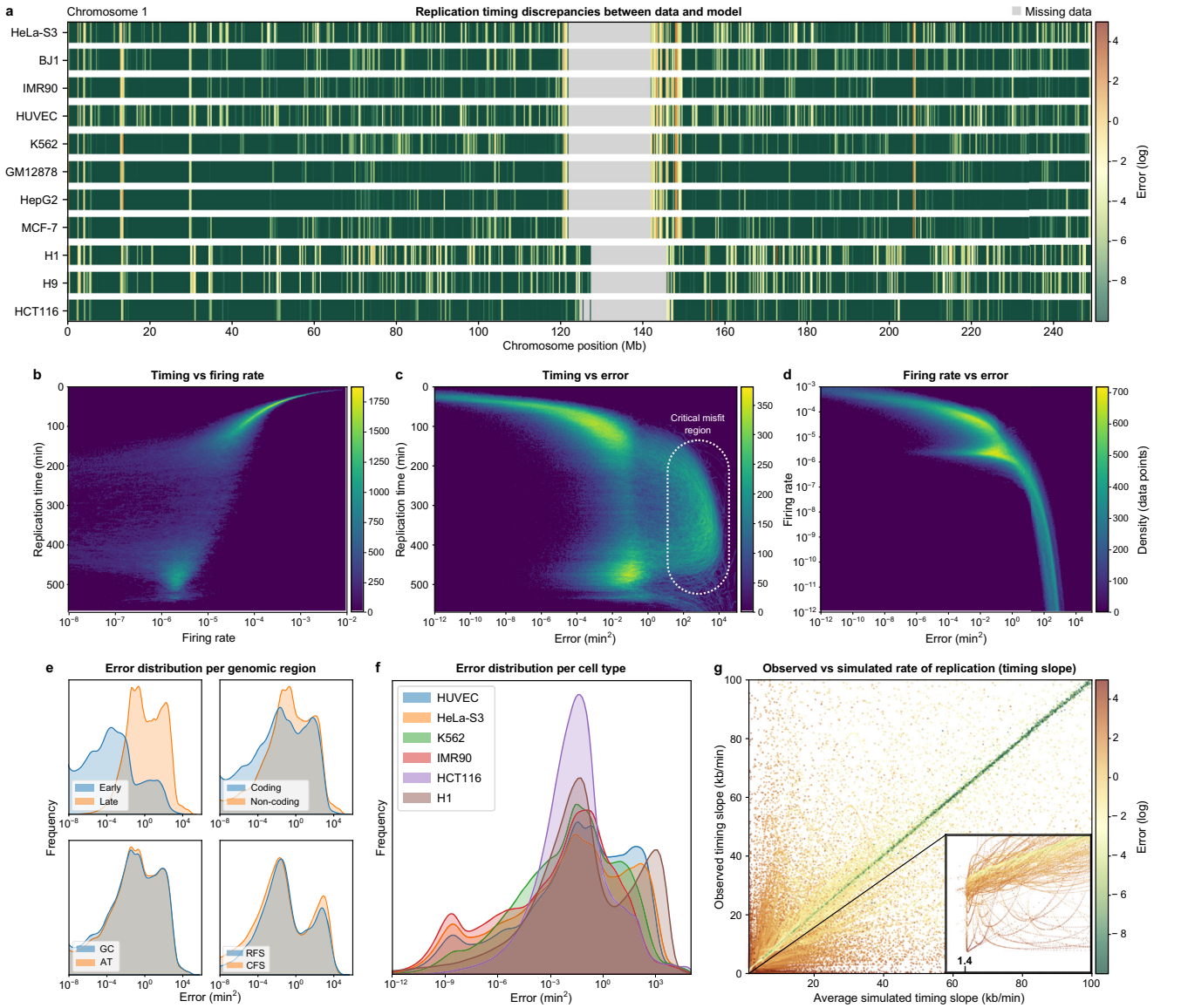



**Fig. 3 | Detecting discrepancies in replication timing determined experimentally and in simulations. a** Normalised error plots (red—high error, green—low error) highlighting deviations between simulated and experimental replication timings (chromosome 1 in various human cell lines). Grey areas: missing or unavailable data. **b**–**d** Density scatter plots illustrating key relationships in H1 cells (averages of 500 simulations). Pairwise combinations of three variables are shown: replication time, firing rate, and error. In **b**, the inverse correlation between replication timing and firing rate is evident, with greater variability in firing rates late in S phase. **c** shows the relationship between replication timing and error, revealing 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, grouped by replication timing (early vs. late), genic vs. intergenic regions, GC vs. AT content, and classification of fragile sites (common vs. rare, CFS vs. RFS). **f** Genome-wide error profiles in different cells. **g** Scatter plot comparing average simulated timing slope, indicative of the progress of replication over time, against observed data, colour-coded by associated error. The zoomed-in region at [1.2, 2] × [0, 2] kb/min highlights the 1.4 kb/min bound on the simulated slope. Each dot represents a simulated-observed data pair, with the strand-like continuity arising from the high resolution of our 1 kb model, where proximity between adjacent pairs reflects the minimal positional shifts captured at this scale.

lengths and active fork numbers, highlights 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 (Fig. 3a illustrates those 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[32]. Replication timing and firing rates show a strong negative correlation (Spearman's rank correlation of ~ −0.89; Fig. 3b); 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 mid-to-late S phase (Fig. 3c). This suggests misfits increase as S phase progresses and fewer firing events occur. Low firing rates are also associated with high errors (Fig. 3d; note the branched profile, reflecting difficulties in accurately modelling high-to-low firing rate transitions). Timing misfits are

https://doi.org/10.1038/s41467-025-59991-w

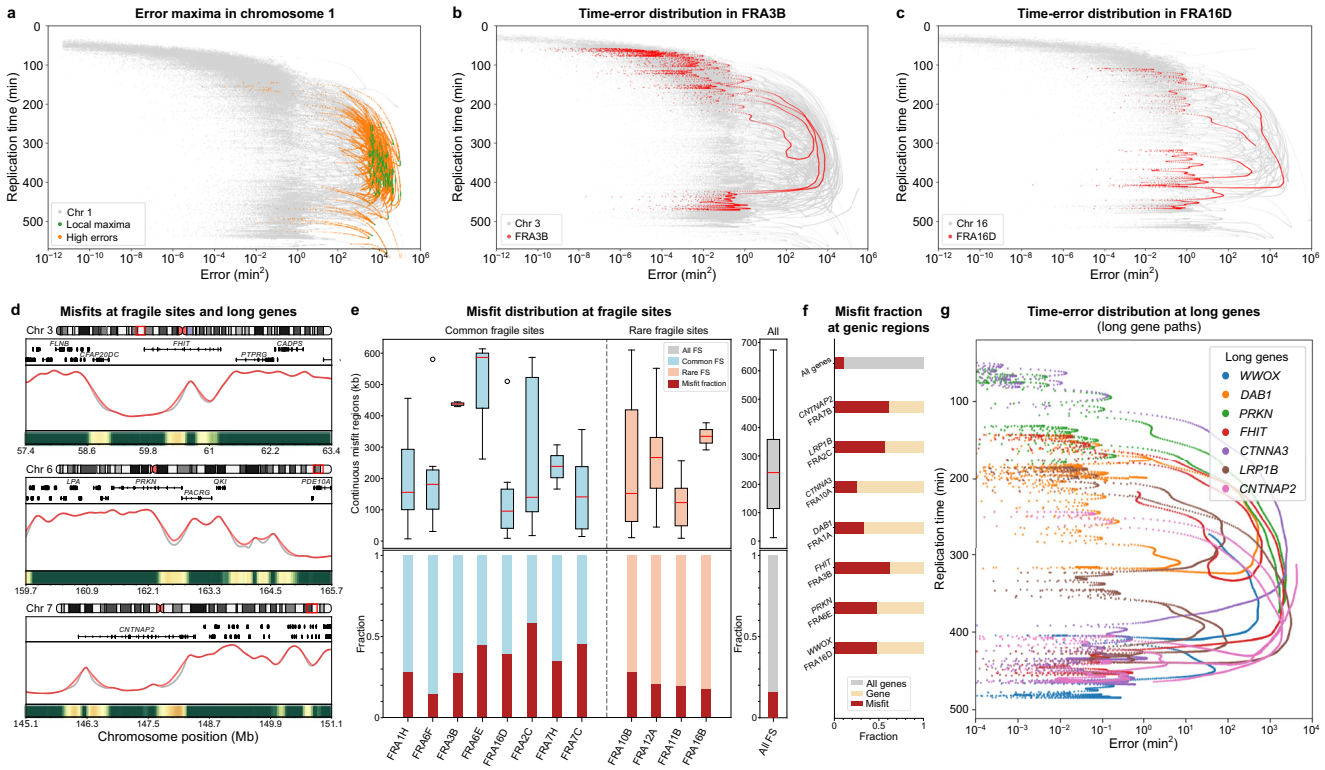



**Fig. 4 | Timing errors in fragile sites and long genes. a** Replication timing vs. error on chromosome 1 in H1, highlighting regions with local maxima in error and neighbouring high-error zones (within a 300 kb radius). The threshold for identifying local maxima in errors is set at $10^{2.8}$ ($min^2$). Each dot represents an error-timing data pair, with the strand-like continuity arising from the high-resolution of our 1 kb model. **b**, **c** Genome-wide scatter plots displaying replication timing vs. error, with specific focus on common fragile sites FRA3B and FRA16D, revealing a continuous error path in mid-to-late replication, near the *FHIT* and *WWOX* genes, respectively. **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, misfit regions overlap with long genes such as *FHIT* (Chr 3), *PRKN* (Chr 6), and *CNTNAP2* (Chr 7). **e** Misfit distribution for common (blue) and rare (pink) fragile sites, compared with the total fragile site misfit fraction (grey). Top: length (in Mb) of continuous misfit regions. Box plots show the median (middle line), 25th–75th percentiles (box), whiskers up to 1.5 times the interquartile range, and outliers (open circles). In total, 1262 continuous misfit regions across all fragile sites were analysed to illustrate global trends. Bottom: normalised misfit fraction at different sites. **f** Misfit fraction analysis of whole-genome genic regions, and at the largest genes within fragile sites (normalised). **g** Scatter plot of replication timing vs. error trajectories for long genes, highlighting error accumulations based on gene size and location within fragile sites.

predominantly concentrated in late-replicating regions (Fig. 3e). This is consistent with prior results suggesting that the replication machinery encounters more obstacles towards the end of S phase[33,34]. 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 HCT116 displaying a distinct pattern likely due to differences in data processing (Fig. 3f; see Methods). Similar disparities were observed in firing rate distributions, hinting at the potential for cell line-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 changes over time—is primarily governed by fork speed, establishing an effective bound of 1.4 kb/min (Fig. 3g). This constraint becomes most evident in regions where observed gradients fall beyond such a 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 long genes

Fragile sites are specific chromosomal regions prone to gap formation or breakage under conditions of replication stress[35]; examples include FRA3B[36] and FRA16D[37]. They frequently arise after inhibiting DNA synthesis or applying other replication stresses[38], often contain few origins[11], and likely result from fork stalling or collapse[39]. Fragile sites can be broadly categorized into common fragile sites (CFSs) present in most of the population and rarer ones (RFSs) found in relatively few individuals[21,40]. As seen before, replication timing misfits tend to be most pronounced in mid-to-late S phase, where regions such as fragile sites often coincide with higher errors (Fig. 4a–d). FRA3B and FRA16D show even greater median misfit lengths (Fig. 4e), indicating that these loci can pose particular challenges for our model. Likewise, large genes in fragile sites (e.g., *CNTNAP2*, *LRP1B*, *FHIT*) exhibit substantial errors (Fig. 4f, g), and applying an error threshold confirms that long genes overlapping misfit regions (~30% of which reside in fragile sites) stand out readily (Supplementary Table 1). Although certain chromosomes (e.g., 15, 20, 22) lack major fragile-site misfits, most display a notable linkage between fragile sites, large genes, and replication stress. While not all fragile sites follow one uniform pattern, these observations underscore broad genomic regulation factors that potentially influence replication timing and error.

Fragile sites are also known to be cell-type specific[36,41], yet our analysis of 11 lines, with H1 cells serving as one illustration, suggests that core fragility-misfit correlations are relatively consistent even if the degree of disruption varies. A more detailed case study of HCT116, where confirmatory data on fragile site expression is available[42], is

https://doi.org/10.1038/s41467-025-59991-w

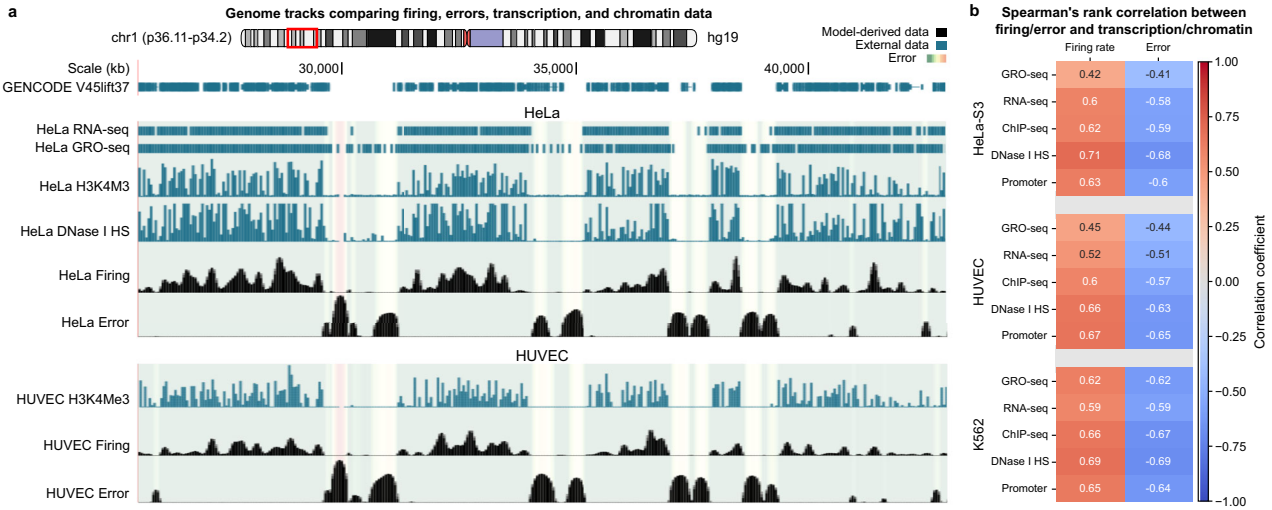



**Fig. 5 | Replication timing discrepancies and firing rate profiles correlate with transcriptional and chromatin data. a** Snapshot from the UCSC Genome Browser showing a detailed view of chromosome 1 (p36.11-p34.2) in HUVEC and HeLa (hg19). Various tracks compare transcriptional and chromatin data to misfit magnitude (error) and firing rate profiles obtained from our model (log-scale). Tracks include RNA-seq (marking mature mRNA levels), GRO-seq (nascent RNA), ChIP-seq for H3K4Me3 (promoters), and DNase I hypersensitivity (open chromatin). The error for each line is represented as a translucent heat map across tracks, with colours ranging from green (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. All tests (two-sided) returned $p$-value $< 10^{-15}$.

included in the Supplementary Information (Supplementary Note 2.4), highlighting how individual chromosomes and cell-line dependent features can shape replication stress at fragile sites. Taken together, these findings reinforce that fragile sites often correlate strongly with replication misfits, though not uniformly across the genome or in every cell type. By pinpointing likely hotspots of replication stress, our model provides a powerful framework for guiding experimental follow-up.

### Transcription and chromatin state

Transcription and replication have long been recognised to interact in complex and sometimes conflicting ways, particularly at fragile sites[43]. Previous studies show that transcription can create barriers to replication, mainly via R-loops, that can obstruct fork progression, leading to stalling or collapse. Long genes associated with CFSs have a scarcity of replication origins, forcing forks to traverse a long distance which can delay replication[44]. This delay is particularly pronounced in transcriptionally active regions. However, this is not always the case, as chromatin structure can play a more dominant role in timing discrepancies. Building on the previous results, we now turn our attention to interactions between transcription, chromatin structure, and replication. Regulatory elements like active promoters and enhancers are marked by histone modifications such as H3K4me3[45], DNase I hypersensitivity (DHS), and transcription-factor binding, detected using ChIP-seq[46,47]. By integrating data from ChIP-seq, RNA-seq, and GRO-seq[18,20,46,48], we assess how these markers are associated with replication timing. Regions with high GRO-seq signals align with peaks in H3K4me3 and DHS signals; they exhibit lower timing errors and higher firing rates (Fig. 5a). Spearman rank correlation analyses reveal varying degrees of association between variables (Fig. 5b). 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 (Supplementary Note 2.5). The consistently higher Spearman rank correlations indicate a strong monotonic relationship, particularly between DHS sites and firing rates, as well as between promoters and firing rates, revealing how chromatin accessibility facilitates replication initiation, even amid non-linear interactions. We observed a moderate to strong negative correlation between GRO-seq and misfits across all 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 state, reducing mechanical barriers to fork progression and lowering the chances of replication stress. Moreover, transcription factor-binding sites have been shown to enhance DNA replication, as evidenced by studies demonstrating that these sites significantly increase replication efficiency[15].

Furthermore, origin density strongly correlates with promoter density[13]. 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 origins in mammalian cells. The strong correlation between high origin firing rates and regions of active transcription, open chromatin, and promoters provides further insight into genome-wide coordination of replication and transcription. Notably, putative origins are often located in open and early-replicating chromatin[17,49] that is well fitted by our model. This synchronisation between replication and transcription may prevent replication stress, particularly during late S phase, aligning with the observation that transcriptionally active euchromatin tends to replicate early, and silent heterochromatin late[50]. 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[51].

## Discussion

In genome-wide simulations, our 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 observations clearly deviate from simulations. While misfit distributions varied across different chromosomes and cell lines (Fig. 2), late-replicating regions consistently exhibited higher misfit rates (Fig. 3). This matches previous findings suggesting these regions are more prone to replication challenges. Firing rates were also strongly

negatively correlated with timing misfits; regions with infrequent origin firing are more susceptible to timing deviations. Additionally, non-coding regions had a higher frequency of misfits, highlighting their potential vulnerability.

We found that many replication-timing misfits occur in proximity to fragile sites and long genes (Fig. 4). Our analysis pools data from multiple studies and cell types[21]. While this provides a broad overview, it does not account for the cell type-specific replication programmes that underlie fragile site expression. Fragile sites are influenced by transcriptional activity and replication timing, both of which vary between cell types[41,52]. For instance, fibroblasts and lymphoblastoid cells exhibit distinct replication initiation patterns, which affect the timing and extent of fragility[36]. We observed consistent trends in the correlation between fragility and replication timing misfits across all 11 lines analysed, with H1 cells used as an illustrative example. This consistency highlights the robustness of the approach in identifying conserved replication dynamics and suggesting candidate regions and genes of interest.

We also performed a statistical assessment of misfit distributions in HCT116, a cell line with robust confirmatory data on fragile site expression[42]. Our results indicate that although fragile sites in HCT116 frequently show statistically significant differences in replication misfits, especially on certain chromosomes, this pattern is not uniform across the entire genome. Further targeted studies could help clarify how cell-line-specific factors shape localized vulnerability within a broader replication error landscape. At the same time, by pinpointing potential replication stress hotspots, our model provides a valuable foundation for deeper experimental investigations into the molecular underpinnings of fragility. Researchers with access to cell type-matched Repli-seq and fragility data could refine this framework to achieve more specific insights.

Additionally, we note that early-replicating fragile sites (ERFS), often linked to highly transcribed genes[53], represent another compelling avenue for future work. For instance, the study by Tubbs et al.[54] demonstrates that poly(dA:dT) tracts can precipitate replication fork collapse in both early- and late-replicating domains, suggesting that similar sequence features may underlie fragility across diverse replication windows. However, robust, high-resolution data on ERFS in human cell lines remain limited; most existing datasets come from mouse[53] or avian cells[55], and cross-species mapping (e.g., via LiftOver) does not reliably capture species-specific replication landscapes. Once comprehensive human datasets—ideally providing cell-type-specific replication timing and validated ERFS coordinates—become available, our model stands ready as a practical tool to assess how misfits at ERFS compare to CFSs and other fragile regions.

Although the model does not incorporate detailed molecular mechanisms, regions with high origin firing rates were nonetheless strongly associated with active transcription, open chromatin, and promoter activity (Fig. 5). These findings align with established knowledge, validating the model and underscoring its robustness. Notably, many misfit regions overlap with known fragile sites or distinct genomic locations, leading to the hypothesis that the model can refine the definition of fragile sites, distinguishing smaller, more nuanced regions of fragility, or even identifying sites prone to replication stress. Such predictions highlight the model's utility in uncovering unexplored genomic vulnerabilities, warranting further experimental validation.

Our approach has various limitations. For instance, we assume that each origin fires independently of others, which may not capture the full complexity of origin licensing and activation (see Methods). However, this simplification allows the model to fit human Repli-seq data rapidly, making it a practical tool for genome-wide analyses. Even so, in reality, a multiplicity of factors (e.g., ORC, Cdc6, and MCM proteins) regulate complex pathways of origin licensing, while later checkpoints and stress response pathways influence cell-cycle progression[56]. Another limitation is that we take no account of higher-order genome structure, but could incorporate data from, for example, Hi-C[6] and the position of R-loops, hairpins and G-quadruplexes that are known to obstruct replication[51]. Furthermore, our model could highlight the relationship between origins and DNA break clusters, such as those found at timing transition regions, which are prone to replication-transcription conflicts and genome instability[57]. Additionally, because Repli-seq data represent population averages, our analysis does not capture potential heterogeneity at the single-cell level[58,59]. Future studies employing single-cell data could thus provide finer resolution of replication dynamics.

Another nuance is that, while our model is quite universal in its assumptions, applying it to organisms like *S. cerevisiae* (budding yeast), which have smaller genomes and precisely located origins[60], may require adjusted parameterisation. In particular, the radius of influence $R$ becomes more critical in a smaller genome where it can have a proportionally greater effect. Outside of autonomously replicating sequences (ARS), the model is expected to assign very low firing rates by default. In Supplementary Note 2.6, we demonstrate the application of our framework to yeast, where the model successfully recovers >86% of known origin locations using only timing data, supporting our hypothesis and highlighting the model's general applicability across eukaryotes. Further investigation in yeast will be presented in a future study.

An exciting application of the model involves exploring the impact of chemotherapies on replication dynamics, particularly those therapies that target the Replication Stress Response (RSR) pathway and its key signalling proteins. By simulating the inhibition of these proteins, the model could provide valuable insights into how these disruptions affect replication timing, origin firing, and potential cell death[61]. This could facilitate prediction of which combination chemotherapies might provide cost-effective approaches to optimise cancer treatments.

## Methods

### Modelling assumptions

Our model is built on several key assumptions. First, the firing time of an origin is modelled as an exponentially distributed random variable, independent of fork movement and of the firing times of other origins. Second, replication forks progress at a constant speed, regardless of the dynamics of origin firing. This constant speed assumption serves as a critical constraint when benchmarking wild-type replication. If fork speed were allowed to vary freely in space and time, it would be possible to adjust fork progression locally to match steep timing profiles, resulting in multiple equally valid solutions and reducing the model's predictive power.

The origin firing rate 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 predictions. 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, our fitting algorithm can effectively turn off potential origins by assigning them a suitably low firing rate.

We also intentionally omit finer details of strand synthesis. In particular, we do not distinguish between leading and lagging strands, nor do we model the formation and joining of Okazaki fragments. By concentrating on the fundamental kinetics driving replication, we gain a clearer understanding of how origin firing rates shape replication timing without introducing unnecessary complexity.

We now present the main framework leading to Eq. (1).

### 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 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

$$T = \min_i \{A_i + |i|/v\} \tag{2}$$

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. (2) is greater than some $t$ if all terms are, which occurs with probability

$$P(T > t) = \prod_i \min\{1, \exp(-f(t - |i|/v))\} \tag{3}$$

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

$$\mathbb{E}[T; n] = \int_0^\infty \prod_i \min\{1, \exp(-f(t - |i|/v))\}\, dt \tag{4}$$

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, 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 1.1 for an explicit 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

$$\mathbb{E}[T; \infty] \equiv \frac{1}{f} \sum_{k=0}^{\infty} \frac{e^{-fk^2/v} - e^{-f(k+1)^2/v}}{2k+1}. \tag{5}$$

With $v$ = 1.4 kb/min[31], Fig. 6a shows the dynamics of $\mathbb{E}[T; n]$ for increasing values of $n$. By relating Eq. (5) to the family of theta and Dawson functions, the following approximation holds (see Supplementary Note 1.2 for a detailed proof)

$$\mathbb{E}[T; \infty] \simeq \frac{1}{2}\sqrt{\frac{\pi}{fv}}. \tag{6}$$

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

$$f_j \simeq \frac{\pi}{4v} T_j^{-2} \tag{7}$$

which provides a first estimate for the intrinsic firing rate of an origin, given its time of replication. Note that Eq.(7) 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[62]. 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. (2) should include the site-specific indexation, for $1 \le j \le n$, as follows

$$T_j = \min_i \{A_i + |i - j|/v\} \tag{8}$$

with indexes congruent modulo $n$, that is, $|i - j| \in \mathbb{Z}/n\mathbb{Z}$ (see Supplementary Note 1). Following a similar argument, the general expression for $\mathbb{E}[T_j; \infty]$, with general firing rates $\{f_i\}$, is given by

$$\mathbb{E}[T_j; \infty] = \sum_{k=0}^{\infty} \frac{e^{-\sum_{|i| \le k}(k-|i|)f_{j+i}/v} - e^{-\sum_{|i| \le k}(k+1-|i|)f_{j+i}/v}}{\sum_{|i| \le k} f_{j+i}}. \tag{9}$$

When $f_j = f$, $\forall j$, Eq. (9) is reduced to Eq. (5). While Eq. (9) 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 $R$, leading to Eq. (1). This parameter represents the radius of replication influence: the distance within which neighbouring origins $\{j - R, \ldots, j - 1, j + 1, \ldots, j + R\}$ 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. (9) should mimic the average 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 $\{f_j\}_{1 \le j \le n}$ as a function of the expectation of $T_j$. Our goal is then to find a solution to Eq. (9), 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 previously explored by Jun, Bechhoefer, and Rhind[22,23,63], revealing other relevant replication metrics, such as inter-origin distances[64]. Our formulation extends these approaches by estimating origin firing rates from discrete replication timing data across the entire human genome, which is discussed next.

### Replication timing data

Replication timing data were sourced and processed from two key databases: the Encyclopedia of DNA Elements (ENCODE[65,66]) and high-resolution Repli-seq from Zhao et al.[28]. 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), HCT116 (colorectal carcinoma line), plus H1 and H9 (embryonic stem cell lines). Data for HUVEC, HeLa, BJ, IMR90, K562, GM12878, HepG2, and MCF-7 cells were obtained from the ENCODE database using the GRCh37 (hg19) human genome assembly[65,66], while data for HCT116, H1, and H9 cells were sourced from high-resolution Repli-seq, using the GRCh38 (hg38) assembly[28].

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

https://doi.org/10.1038/s41467-025-59991-w

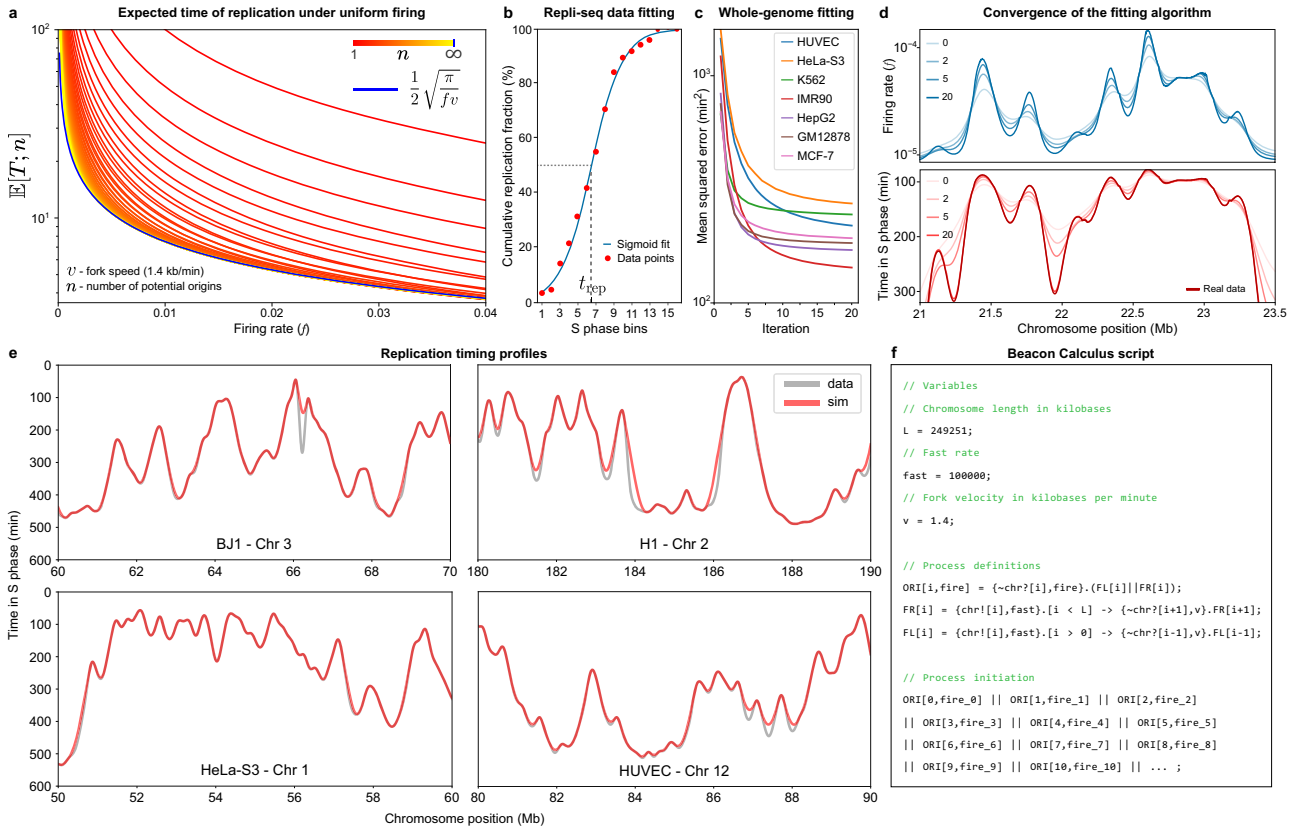



**Fig. 6 | Fitting the model. a** Replication asymptotics under uniform firing: logarithmic plot of the expected replication time, $\mathbb{E}[T;n]$, as a function of the firing rate, $f$, and the number of potential origins, $n$ (spaced at 1 kb intervals), for $1 \le n < \infty$, with $v = 1.4$ kb/min. 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 the median replication time, $t_{rep}$ (the instant 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 ~3 min 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. (7), and the corresponding timing profile (below). **e** Observed (Repli-seq) timing against the simulated profiles for different lines and genomic regions. **f** Model written in the Beacon Calculus process algebra. 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. (1).

of replication across the genome[67]. 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. 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.[28]. We consider the median replication time, $t_{rep}$, defined as the bin value $t$ where $F_{rep}(t) = 50\%$, indicating that half of the cell population has completed replication (Fig. 6b). Although Eq. (9) 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[68], we do not incorporate them into our analysis. Repli-seq data shows consistent patterns across different cell lines. We present representative results from multiple lines, but specific analyses may be more suitable for certain cases, depending on the availability and quality of the data. Although regions with repetitive sequences or low complexity are often poorly mapped using Repli-seq data[28,65], these regions account for ~ 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 (Supplementary Note 2.3).

## Fitting algorithm

Equation (9) establishes a continuous, monotonic relationship between each firing rate, $f_j$, and its corresponding replication time, $\tilde{T}_j$. Our aim is to infer the set of firing rates $\{f_j\}$ from experimentally measured timing data $\{T_j\}$. Rather than relying on large-scale simulations, we employ an iterative procedure that leverages the monotonic relationship in Eq. (9): each firing rate is updated so as to minimize the difference between $\tilde{T}_j$ (predicted via Eq. (9)) and $T_j$ (obtained from Repli-seq data). In practice, for large $n$, Eq. (9) provides a good approximation of $\tilde{T}_j$. We initialize each site $j$ by

$$f_j(0) = \frac{\pi}{4v} T_j^{-2}, \tag{10}$$

then iteratively refine its firing rate according to

$$f_j(k+1) = f_j(k)\left(\frac{\tilde{T}_j(k)}{T_j}\right)^{\alpha}, \tag{11}$$

where $\tilde{T}_j(k)$ is the predicted replication time at iteration $k$, and $T_j$ is the experimentally observed timing. The exponent $\alpha$ governs how strongly firing rates respond to each site's misfit, behaving much like a fixed-point iteration or an inexact gradient descent. Numerical experiments suggest $\alpha = 2$ provides a robust balance between speed of convergence and numerical stability, but other choices of $\alpha$ are feasible if the data require finer control or if a gradient-based approach is desired. Every 1 kb segment is treated initially as a potential origin, but the algorithm's updates naturally drive most firing rates to negligible values, reflecting the selective activation of origins in the genome. The radius of neighbouring influence, $R$, may be refined for optimisation. We track convergence by measuring each site's fit error, defined as the squared difference $(T_j - \tilde{T}_j(k))^2$, in min$^2$. Because the method directly leverages the convolution-like form of Eq. (9), it avoids repeating large-scale simulations. This efficiency means that fitting 3.2 million sites per human genome can typically be done within a few minutes on a single Intel Ice Lake CPU (Fig. 6c–e). Thus, although we rely on simple iterative corrections, the monotonic structure of $\tilde{T}_j$ with respect to $f_j$ ensures the scheme converges reliably, provided $\alpha$ and other parameters remain moderate. We have added a more detailed discussion of this algorithm, including its convolution interpretation, in Supplementary Note 2.2.1.

### Simulations

To simulate replication, we use Beacon Calculus (`bcs`)[27], a process algebra designed for simulating biological systems. In this framework, each component of the system is treated as a process capable of executing certain actions, each governed by an exponential rate. The simulation uses a modified Gillespie algorithm, allowing multiple processes to run in parallel, a property that is especially important for modelling DNA replication, where many events occur concurrently. In our case, we represent replication with three processes: replication origin firing (`ORI`), and passive replication by left- (`FL`), and right-moving forks (`FR`). Each process is associated with a specific position at a given resolution on a chromosome of length `L`, and origins have an additional parameter, the firing rate, `fire`, or $f$ in our model (Fig. 6f).

In each `bcs` simulation, when a process is activated and its associated action is executed (i.e., when a random variable is realised), the time and location of that event is recorded. Our model operates at a 1 kb resolution, so each action is assigned to a specific 1 kb segment; consequently, the firing of an origin (`ORI`) or the passive replication by a fork (`FL` or `FR`) is registered as an event within that segment. All sites are treated as potential origins; however, the fitting algorithm effectively turns them on or off by assigning them high or near-zero firing rates, respectively. The model is flexible and can be adapted to different resolutions. In our current implementation, an origin is defined as corresponding to a 1 kb segment. This computational implementation is independent from the analytical approach described above, providing a means to confirm the closed-form mathematical analysis and explore additional replication features. Further details on the `bcs` formalism and its usage are discussed in Supplementary Note 2.1.

In `bcs`, $\nu$ is treated as the constant replication rate of a moving fork, i.e., the parameter of an exponential distribution governing the time required to passively replicate one site. This differs from the constant fork speed assumption underlying Eq. (9). Specifically, in the `bcs` case, the time $F_k$ required for a fork to replicate $k$ consecutive sites follows an Erlang($k$, $\nu$) distribution, meaning that $\mathbb{E}[F_{|i-j|}] = |i-j|/\nu$, which mirrors the approximation used in Eq. (8). Therefore, when averaged over a sufficiently large number of simulations, 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[31], and results were averaged over 500 simulations, with the radius of influence set to $R$ = 2000 kb, as previously defined.

### Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The replication timing data used in this study were obtained from the Encyclopedia of DNA Elements (ENCODE) using the GRCh37 (hg19) assembly[65,66] and from high-resolution Repli-seq aligned to the GRCh38 (hg38) assembly[28]. RNA-seq[18], ChIP-seq[19], and GRO-seq[20] datasets were also obtained from ENCODE. Additionally, GRO-seq data used in this study were accessed from the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession numbers GSE62046, GSE94872, and GSE60454. Fragile site locations were sourced from the publicly available HumCFS database[21], accessible at https://webs.iiitd.edu.in/raghava/humcfs/. All these data are publicly available.

## Code availability

The source code implementing the main fitting algorithm, together with the replication timing fit error and origin firing rate bedgraph files, is available at: https://github.com/fberkemeier/DNA_replication_model.git (version 1.0.0, https://doi.org/10.5281/zenodo.15337522)[69]. Beacon Calculus simulations were carried out using version 1.1.0 of `bcs`, accessible at https://github.com/MBoemo/bcs[27]. Supplementary Note 2 provides additional examples of `bcs` scripts and optimisation algorithms.

## References


1. Gefter, M. L. DNA replication. *Annu. Rev. Biochem.* **44**, 45–78 (1975).
2. Leonard, A. C. & Méchali, M. DNA replication origins. *Cold Spring Harb. Perspect. Biol.* **5**, a010116 (2013).
3. Waga, S. & Stillman, B. The DNA replication fork in eukaryotic cells. *Annu. Rev. Biochem.* **67**, 721–751 (1998).
4. Mirkin, E. V. & Mirkin, S. M. Replication fork stalling at natural impediments. *Microbiol. Mol. Biol. Rev.* **71**, 13–35 (2007).
5. Rhind, N. & Gilbert, D. M. DNA replication timing. *Cold Spring Harb. Perspect. Biol.* **5**, a010132 (2013).
6. Marchal, C., Sima, J. & Gilbert, D. M. Control of DNA replication timing in the 3D genome. *Nat. Rev. Mol. Cell Biol.* **20**, 721–737 (2019).
7. Müller, C. A. & Nieduszynski, C. A. DNA replication timing influences gene expression level. *J. Cell Biol.* **216**, 1907–1914 (2017).
8. Briu, L.-M., Maric, C. & Cadoret, J.-C. Replication stress, genomic instability, and replication timing: a complex relationship. *Int. J. Mol. Sci.* **22**, 4764 (2021).
9. Koren, A. et al. Genetic variation in human DNA replication timing. *Cell* **159**, 1015–1026 (2014).
10. Woo, Y. H. & Li, W.-H. DNA replication timing and selection shape the landscape of nucleotide variation in cancer genomes. *Nat. Commun.* **3**, 1004 (2012).
11. Sinai, M. I. T. & Kerem, B. DNA replication stress drives fragile site instability. *Mutat. Res. Fundam. Mol. Mech. Mutagen.* **808**, 56–61 (2018).
12. Smith, D. I., Zhu, Y., McAvoy, S. & Kuhn, R. Common fragile sites, extremely large genes, neural development and cancer. *Cancer Lett.* **232**, 48–57 (2006).
13. Sequeira-Mendes, J. et al. Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet.* **5**, e1000446 (2009).
14. Ehrenhofer-Murray, A. E. Chromatin dynamics at DNA replication, transcription and repair. *Eur. J. Biochem.* **271**, 2335–2349 (2004).

15. Turner, W. J. & Woodworth, M. E. DNA replication efficiency depends on transcription factor-binding sites. *J. Virol.* **75**, 5638–5645 (2001).
16. Guilbaud, G. et al. Evidence for sequential and increasing activation of replication origins along replication timing gradients in the human genome. *PLoS Comput. Biol.* **7**, e1002322 (2011).
17. Audit, B. et al. Open chromatin encoded in DNA sequence is the signature of 'master' replication origins in human cells. *Nucleic Acids Res.* **37**, 6064–6075 (2009).
18. Marguerat, S. & Bähler, J. RNA-seq: from technology to biology. *Cell. Mol. Life Sci.* **67**, 569–579 (2010).
19. Pepke, S., Wold, B. & Mortazavi, A. Computation for ChIP-seq and RNA-seq studies. *Nat. Methods* **6**, S22–S32 (2009).
20. Lopes, R., Agami, R. & Korkmaz, G. GRO-seq, a tool for identification of transcripts regulating gene expression. *Promoter Associated RNA Methods Protocols*, 45–55 (2017).
21. Kumar, R. et al. HumCFS: a database of fragile sites in human chromosomes. *BMC Genomics* **19**, 1–8 (2019).
22. Jun, S., Zhang, H. & Bechhoefer, J. Nucleation and growth in one dimension. I. The generalized Kolmogorov–Johnson–Mehl–Avrami model. *Phys. Rev. E* **71**, 011908 (2005).
23. Jun, S. & Bechhoefer, J. Nucleation and growth in one dimension. II. Application to DNA replication kinetics. *Phys. Rev. E* **71**, 011909 (2005).
24. de Moura, A. P. S., Retkute, R., Hawkins, M. & Nieduszynski, C. A. Mathematical modelling of whole chromosome replication. *Nucleic Acids Res.* **38**, 5623–5633 (2010).
25. Retkute, R., Nieduszynski, C. A. & de Moura, A. P. S. Mathematical modeling of genome replication. *Phys. Rev. E* **86**, 031916 (2012).
26. Arbona, J.-M. et al. Neural network and kinetic modelling of human genome replication reveal replication origin locations and strengths. *PLoS Comput. Biol.* **19**, e1011138 (2023).
27. Boemo, M. A., Cardelli, L. & Nieduszynski, C. A. The Beacon Calculus: a formal method for the flexible and concise modelling of biological systems. *PLoS Comput. Biol.* **16**, e1007651 (2020).
28. Zhao, P. A., Sasaki, T. & Gilbert, D. M. High-resolution Repli-seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biol.* **21**, 1–20 (2020).
29. Anderson, C. J. et al. Strand-resolved mutagenicity of DNA damage and repair. *Nature*. **630**, 744–751 (2024).
30. Técher, H. et al. Replication dynamics: biases and robustness of DNA fiber analysis. *J. Mol. Biol.* **425**, 4845–4855 (2013).
31. Conti, C. et al. Replication fork velocities at adjacent replication origins are coordinately modified during DNA replication in human cells. *Mol. Biol. Cell* **18**, 3059–3067 (2007).
32. Bracci, A. N. et al. The evolution of the human DNA replication timing program. *Proc. Natl. Acad. Sci. USA* **120**, e2213896120 (2023).
33. Branzei, D. & Foiani, M. Maintaining genome stability at the replication fork. *Nat. Rev. Mol. Cell Biol.* **11**, 208–219 (2010).
34. Colicino-Murbach, E., Hathaway, C. & Dungrawala, H. Replication fork stalling in late S-phase elicits nascent strand degradation by DNA mismatch repair. *Nucleic Acids Res.* **52**, 10999–11013 (2024).
35. Li, S. & Wu, X. Common fragile sites: protection and repair. *Cell Biosci.* **10**, 1–9 (2020).
36. Letessier, A. et al. Cell-type-specific replication initiation programs set fragility of the FRA3B fragile site. *Nature* **470**, 120–123 (2011).
37. Palakodeti, A., Han, Y., Jiang, Y. & Le Beau, M. M. The role of late/slow replication of the FRA16D in common fragile site induction. *Genes, Chromosomes and Cancer* **39**, 71–76 (2004).
38. Glover, T. W., Wilson, T. E. & Arlt, M. F. Fragile sites in cancer: more than meets the eye. *Nat. Rev. Cancer* **17**, 489–501 (2017).
39. Kaushal, S. & Freudenreich, C. H. The role of fork stalling and DNA structures in causing chromosome fragility. *Genes, Chromosomes and Cancer* **58**, 270–283 (2019).
40. Frankish, A. et al. GENCODE: reference annotation for the human and mouse genomes in 2023. *Nucleic Acids Res.* **51**, D942–D949 (2023).
41. Le Tallec, B. et al. Molecular profiling of common fragile sites in human fibroblasts. *Nat. Struct. Mol. Biol.* **18**, 1421–1423 (2011).
42. Boteva, L. et al. Common fragile sites are characterized by faulty condensin loading after replication stress. *Cell Rep.* **32**, 108177 (2020).
43. Knott, S. R. V., Viggiani, C. J. & Aparicio, O. M. To promote and protect: coordinating DNA replication and transcription for genome stability. *Epigenetics* **4**, 362–365 (2009).
44. Blin, M. et al. Transcription-dependent regulation of replication dynamics modulates genome stability. *Nat. Struct. Mol. Biol.* **26**, 58–66 (2019).
45. Huang, X. et al. Stable H3K4me3 is associated with transcription initiation during early embryo development. *Bioinformatics* **35**, 3931–3936 (2019).
46. Cockerill, P. N. Structure and function of active chromatin and DNase I hypersensitive sites. *FEBS J.* **278**, 2182–2210 (2011).
47. Young, M. D. et al. ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic Acids Res.* **39**, 7415–7427 (2011).
48. Crawford, G. E. et al. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc. Natl. Acad. Sci. USA* **101**, 992–997 (2004).
49. Chen, Y.-H. et al. Transcription shapes DNA replication initiation and termination in human cells. *Nat. Struct. Mol. Biol.* **26**, 67–77 (2019).
50. Gilbert, D. M. Replication timing and transcriptional control: beyond cause and effect. *Curr. Opin. Cell Biol.* **14**, 377–383 (2002).
51. García-Muse, T. & Aguilera, A. Transcription–replication conflicts: how they occur and how they are resolved. *Nat. Rev. Mol. Cell Biol.* **17**, 553–563 (2016).
52. Brison, O. et al. Transcription-mediated organization of the replication initiation program across large genes sets common fragile sites genome-wide. *Nat. Commun.* **10**, 5693 (2019).
53. Barlow, J. H. et al. Identification of early replicating fragile sites that contribute to genome instability. *Cell* **152**, 620–632 (2013).
54. Tubbs, A. et al. Dual roles of poly(dA:dT) tracts in replication initiation and fork collapse. *Cell* **174**, 1127–1142 (2018).
55. Pentzold, C. et al. FANCD2 binding identifies conserved fragile sites at large transcribed genes in avian cells. *Nucleic Acids Res.* **46**, 1280–1294 (2018).
56. Boos, D. & Ferreira, P. Origin firing regulations to control genome replication timing. *Genes* **10**, 199 (2019).
57. Corazzi, L. et al. Linear interaction between replication and transcription shapes DNA break dynamics at recurrent DNA break clusters. *Nat. Commun.* **15**, 3594 (2024).
58. Massey, D. J. & Koren, A. High-throughput analysis of single human cells reveals the complex nature of DNA replication timing control. *Nat. Commun.* **13**, 2402 (2022).
59. Dileep, V. & Gilbert, D. M. Single-cell replication profiling to measure stochastic variation in mammalian replication timing. *Nat. Commun.* **9**, 427 (2018).
60. Siow, C. C., Nieduszynska, S. R., Müller, C. A. & Nieduszynski, C. A. OriDB, the DNA replication origin database updated and extended. *Nucleic Acids Res.* **40**, D682–D686 (2012).
61. Manic, G. et al. Replication stress response in cancer stem cells as a target for chemotherapy. *Semin. Cancer Biol.* **53**, 31–41 (2018).
62. Rhind, N., Yang, S. C.-H. & Bechhoefer, J. Reconciling stochastic origin firing with defined replication timing. *Chromosome Res.* **18**, 35–43 (2010).
63. Jun, S. & Rhind, N. Just-in-time DNA replication. *Physics* **1**, 32 (2008).

64. Herrick, J., Jun, S., Bechhoefer, J. & Bensimon, A. Kinetic model of DNA replication in eukaryotic organisms. *J. Mol. Biol.* **320**, 741–750 (2002).
65. Hansen, R. S. et al. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl Acad. Sci.* **107**, 139–144 (2010).
66. Davis, C. A. et al. The encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).
67. Thurman, R. E. et al. Identification of higher-order functional domains in the human ENCODE regions. *Genome Res.* **17**, 917–927 (2007).
68. Massey, D. J. & Koren, A. Telomere-to-telomere human DNA replication timing profiles. *Sci. Rep.* **12**, 9560 (2022).
69. Berkemeier, F., Cook, P. R. & Boemo, M. A. DNA replication timing reveals genome-wide features of transcription and fragility (this paper). *DNA_replication_model* GitHub repository, https://doi.org/10.5281/zenodo.15337522 (2025).
70. Hulke, M. L., Massey, D. J. & Koren, A. Genomic methods for measuring DNA replication dynamics. *Chromosome Res.* **28**, 49–67 (2020).

## Acknowledgements

We thank Prof. Sarah McClelland (Barts Cancer Institute, Queen Mary University of London) and Dr. Mathew Jones (Frazer Institute, University of Queensland) for their 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 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). This work was made possible by the Leverhulme Trust Research Project Grant RPG-2022-028, which was successfully applied for and secured by M.A.B. Additional funding and career support were provided by a Rokos Postdoctoral Associate position at Queens' College Cambridge to F.B. and a fellowship at St John's College, Cambridge to M.A.B.

## Author contributions

The project was conceived by F.B. and M.A.B. F.B. was responsible for conceptualisation, methodology and formal analysis. F.B., M.A.B., and P.R.C. contributed to data interpretation. F.B. was responsible for writing the original draft. M.A.B. and P.R.C. reviewed and edited the manuscript. M.A.B. applied for and was awarded the funding.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-59991-w.

**Correspondence** and requests for materials should be addressed to Francisco Berkemeier or Michael A. Boemo.

**Peer review information** *Nature Communications* thanks Alessandro de Moura, Masatoshi Fujita and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at http://www.nature.com/reprints

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2025

# 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 $|i| \leq vt \leq |i + 1|$. 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

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

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

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

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

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

Solving the integrals yields

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

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

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

In particular,

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

which is Equation (5). Equation (1) 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 Equation (S12) in a more compact, non-dimensional form

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

As $n \to \infty$, we have

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

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[1], 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 Equation (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[2]. 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

$$\mathbb{E}[T; \infty] \simeq \frac{1}{2}\sqrt{\frac{\pi}{fv}} \tag{S21}$$

as in Equation (6).

# Supplementary Note 2: Computational methods and data

## 2.1 Beacon Calculus model

As discussed in Boemo et al.[3], a minimal replication model in `bcs` is built around three core processes: replication origins (`ORI`), left-moving forks (`FL`), and right-moving forks (`FR`). These processes are positioned along a chromosome of length `L`, where each process has a unique integer parameter, `i`, representing its specific location between 1 and `L`. In addition, the origins have a replication initiation rate, `fire` (or $f$ in our model), which can be extended to include licensing probabilities in more advanced setups. To track which positions have been replicated, `bcs` uses markers called beacons. Whenever a fork replicates position `i`, it dispatches a beacon on the `chr` channel with the parameter `i`. This beacon indicates that replication is complete at that coordinate, ensuring the model can monitor progress across the entire chromosome.

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[45,0.0010355526093646997] || 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 the model

### 2.2.1 Main algorithm

The following code presents the main fitting function, `fitfunction`, used in the fitting algorithm described in the main text. It provides an efficient way of computing Equation (1) 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` (radius of influence $R$, as dedfined in the main text). The first guess `x00` is then constructed based on `list`, by Equation (7). We use an adapted version of `np.roll()`. Data was processed via the Python extension pyBigWig[4]. 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)
```

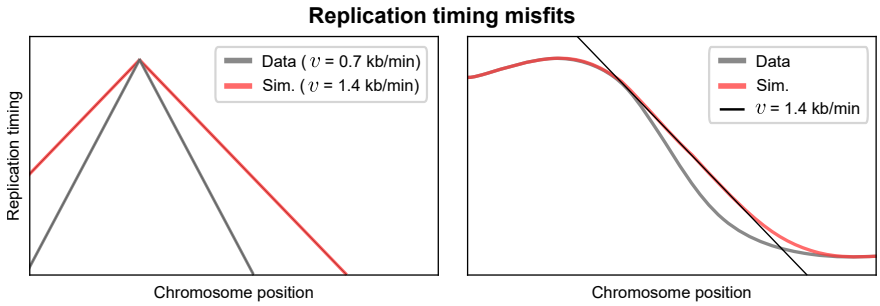
```
            for j in range(len(x0)):
                if fitf(time, lst, x0, j) < 10**(-20):
                    result[j] = 10**(-20)
                elif abs(time[j] - lst[j]) < .5:
                    result[j] = x0[j]
                else:
                    result[j] = fitf(time, lst, x0, j)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

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

### 2.2.2 Effects of fork speed on replication timing misfits

Our model pairs a fixed fork speed with stochastic origin firing, imposing a natural limit on how steeply replication timing can transition from early to late. Even if more origins fire in a region, they can only flatten this slope; they cannot exceed the fork-speed bound. Consequently, any empirical data showing sharper transitions—often due to fork stalling or replication stress—will remain under-fitted (i.e., predicted to replicate too early). While the fitting algorithm may raise firing rates to accommodate steep timing, it does not systematically inflate them; once the constant-speed ceiling is reached, persistent misfits highlight regions where forks slow or stall beyond our model's assumptions.

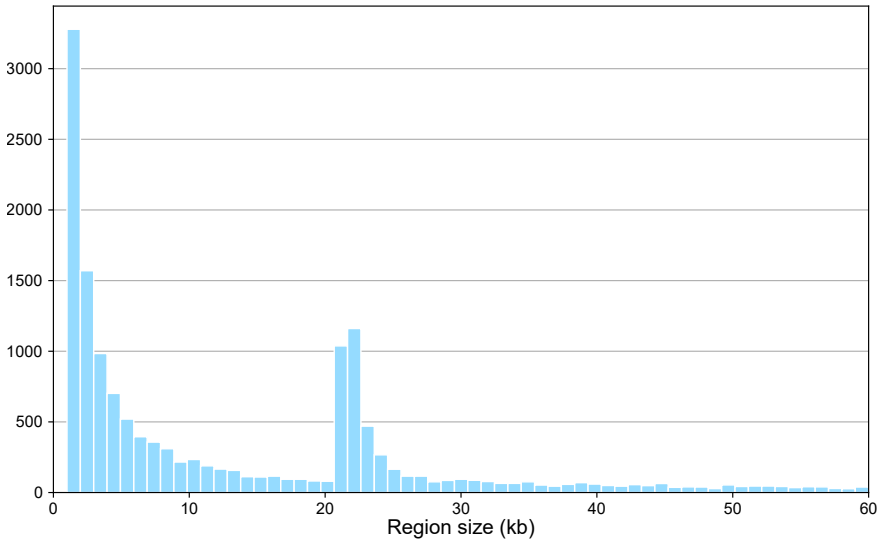
This is most clearly illustrated by thinking about the timing curve of a single-origin system: there will be a sharp point at the origin and the gradient elsewhere will be determined by the rate of fork movement. Adding additional origins at any firing rate can only decrease the magnitude of the curve's gradient.




**Supplementary Figure 1: Effects of fork speed on replication timing profiles.**
Illustration of how a fixed fork speed constrains the steepness of the replication-timing curve. Left: A single-origin "peak" shows that, if the replication fork moves more slowly (gray line, $v = 0.7$ kb/min), the slope is steeper than a faster fixed fork speed can reproduce (red line, $v = 1.4$ kb/min). Allowing extra origins to fire simply flattens slopes rather than raising them above the fork-speed limit. Right: In a multi-origin context, empirical data (gray) can exhibit sharper early-to-late transitions than the model (red) allows. Once fork speed (black diagonal line) is reached, the model cannot replicate any faster, leading to a systematic underestimation of the replication time.

## 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[5,6]. Based on data from Hansen et al.[5], 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 (Supplementary Fig. 2). 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.



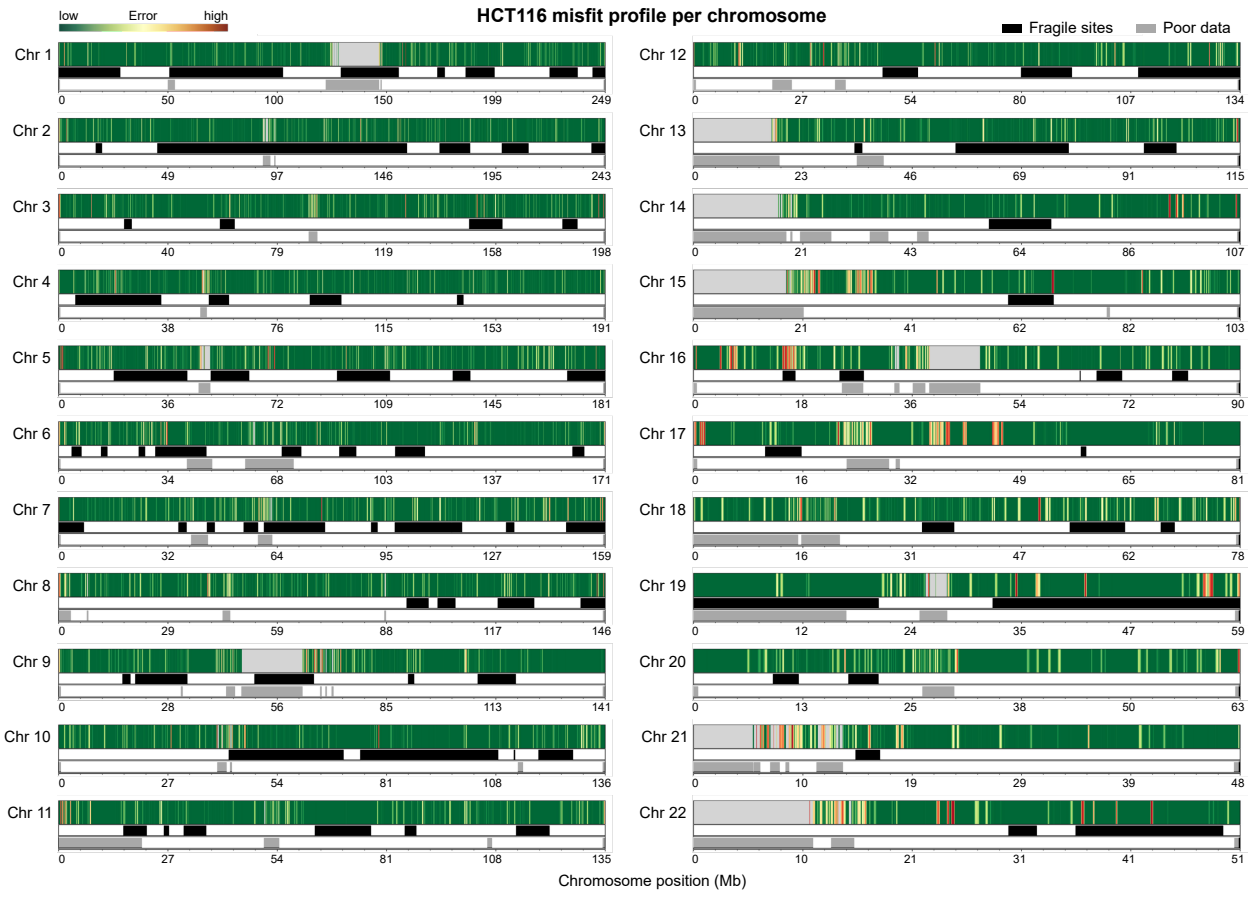

**Supplementary Figure 2: Distribution of problematic mappability region sizes.**
Histogram showing the distribution of region sizes with low or problematic mappability (in kilobases) across the genome. These regions are excluded from replication timing analyses due to difficulties in accurately mapping 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 Fragility analysis in HCT116

This section expands upon the main text's examination of fragile sites and replication-timing misfits, with a focus on whether specific loci in HCT116 display distinctive error patterns. HCT116 was chosen since confirmatory data on fragile site expression is available[6,7]. The analyses presented here include a high-resolution mapping of misfit regions and additional statistical comparisons of fragile versus non-fragile sites. These efforts help clarify whether fragile sites generally stand out from the rest of the genome in terms of replication timing errors, or if notable differences only arise at certain chromosomes or loci.

### 2.4.1 Misfit regions in HCT116

We begin by compiling a chromosome-scale overview of replication misfits in HCT116, focusing on regions where high error levels overlap with known fragile sites. Supplementary Fig. 3 aligns misfit values with fragile sites, making it possible to see whether errors concentrate in these loci or appear at similar levels elsewhere in the genome. In some cases, fragile sites show misfits comparable to surrounding regions. In others, they deviate markedly, suggesting that local or cell-line-specific factors may be at play. An accompanying table (supplementary file '`misfit_genes_HCT116.xlsx`') lists each gene found within high-misfit regions, along with coordinates, length, and fragile-site status. This resource helps identify instances where replication-timing anomalies coincide with genes of particular interest, such as large or transcriptionally active loci. By linking these observations to fragile-site positions, we can pinpoint areas that may warrant further study. All downstream analyses exclude centromeres, telomeres, assembly gaps, and bins with low mappability.

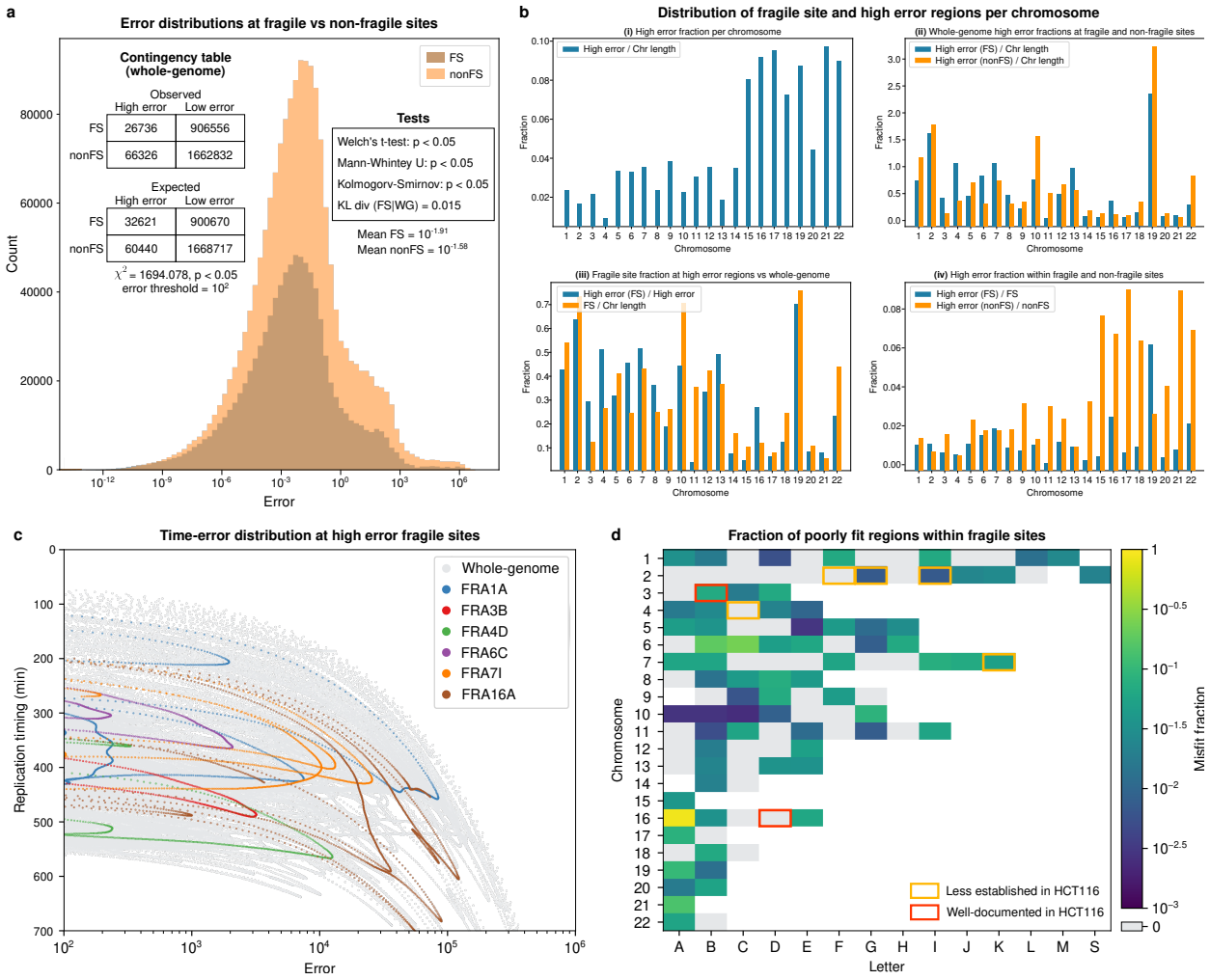



**Supplementary Figure 3: Whole-genome misfit profiles of HCT116 cells.**
Error heatmaps produced by the main model fitted to Repli-seq data for all chromosomes of HCT116 cells. Black bands indicate fragile sites reported in the HumCFS database, whereas gray bands denote poorly fit segments, including centromeres, telomeres, and regions with known mappability issues.

### 2.4.2 Statistical analysis

To investigate whether fragile sites (FS) differ from non-fragile sites (nonFS) in replication error distributions, we carry out two complementary comparisons. First, we examine the entire set of genomic positions, comparing all FS with nonFS without applying any error threshold. Second, we introduce a cutoff of $10^2$ ($\text{min}^2$) to define high-error positions and compare FS and nonFS solely within this subset. By setting the threshold at $10^2$, we capture sufficiently large samples of positions that exceed moderate error values, ensuring both statistical power and biological relevance. This two-tiered approach enables us to assess broad differences in replication error profiles (the full dataset) as well as to focus specifically on positions where replication stress is presumably more pronounced (the high-error subset).

When we consider all genomic positions, FS and nonFS show highly significant differences in both replication error and replication timing. Welch's $t$-tests, Mann–Whitney $U$ tests, and Kolmogorov–Smirnov tests all return near-zero $p$-values (Supplementary Fig. 4a), reflecting sig-




**Supplementary Figure 4: Error statistics in HCT116 fragile sites.**
**a** Error distribution on fragile sites (FS) and non-fragile sites (nonFS). Left of histogram: contingency tables for observed and expected counts of FS/nonFS and high/low errors, and Chi-square results, with an error threshold of $10^2$. Right: statistical tests results between FS and nonFS error distributions. **b** Overview of how replication errors and fragile sites distribute across multiple comparisons: **(i)** The fraction of high-error loci among all loci on the chromosome, **(ii)** the fraction of high-error loci at FS vs. the fraction of high-error loci at nonFS (both normalized by total chromosome length), **(iii)** the fraction of FS among high-error loci vs. the fraction of FS among all loci, and **(iv)** the fraction of high-error FS among all FS vs. the fraction of high-error nonFS among all nonFS. **c** Relication timing vs. error at poorly fitted fragile sites. **d** Heatmap displaying the high error fractions of fragile sites, across all sites reported in HumCFS. Well-documented (FRA3B and FRA16D) and less established (FRA2F, FRA2G, FRA2I, FRA4C, and FRA7K) FS in HCT116 are highlighted.

nificant shifts in their means, medians, and overall shapes. Large negative $t$-values indicate that, on average, FS exhibit lower errors and replicate earlier than nonFS. We also compare the one-dimensional error distribution at FS with that of the genome at large using the Kullback–Leibler divergence $D_{\mathrm{KL}} \approx 0.015$, suggesting a moderate departure between these distributions. Extending this analysis to the two-dimensional time-error distribution yields a slightly higher $D_{\mathrm{KL}} \approx 0.03$, confirming that FS differ from the genome in a joint replication context, yet not to an extreme degree.

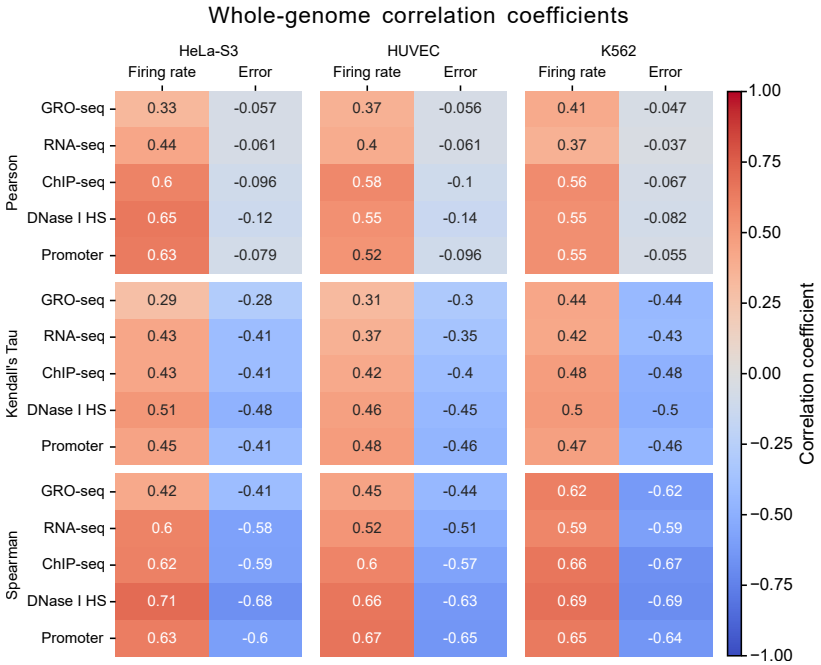
Focusing on regions with errors above $10^2$, we see a strong association between FS status and error classification (Chi-square = 1694.078, $p \approx 0$; contingency table in Supplementary Fig. 4a). Restricting our analysis to these high-error loci, Welch's $t$-test suggests that the mean error is slightly higher at FS ($t = 3.707$, $p = 2.1 \times 10^{-4}$), while replication timing is earlier ($t = -6.172$, $p = 6.78 \times 10^{-10}$). The two-dimensional Hotelling's $T^2$ result ($T^2 = 69.748$, $p = 7.77 \times 10^{-16}$) confirms that FS maintain distinct replication features even within this high-error subset. In other words, FS are somewhat under-represented among positions exceeding the threshold, yet those that do exceed $10^2$ display a characteristic profile of modestly elevated errors and notably earlier replication timing compared to other high-error regions. These findings also point to potential fragile sites in HCT116 that remain understudied, highlighting them as candidates for closer experimental investigation.

At the genome-wide level, these findings suggest that FS often exhibit fewer extreme errors than would be expected by chance. However, more detailed analyses of specific chromosomes, where FS in HCT116 are well documented[8,6], reveal substantial variability in how these errors manifest across different genomic contexts. Supplementary Fig. 4b shows the chromosome-based distributions of errors and FS vs. nonFS, illustrating the FS-dependent nature of replication-stress patterns. Of particular note is the increasing high-error fraction on chromosomes 3 and 16, and the distinct dynamics on chromosome 19. Supplementary Fig. 4c highlights high-error time–error distributions of poorly fit FS, while Supplementary Fig. 4d presents the overall high-error fraction in FS from the HumCFS database[9]. Interestingly, our model flags the rare fragile site FRA16A as especially problematic[10,11] and predicts replication-stress signatures at the well-established FRA3B[12,13]. In contrast, the model appears to fit well at FRA16D, which aligns with studies challenging its instability in HCT116[7]. Our analysis also identifies less established FS in HCT116—FRA2G, FRA2I, and FRA7K[7,14,15]—as potentially of interest.

Overall, these findings show that although fragile sites frequently show statistically significant differences in replication misfits, particularly on certain chromosomes, this pattern does not hold uniformly across the entire genome. Further targeted studies may help untangle how cell-line-specific factors shape these localized vulnerability profiles within broader replication error landscapes. Nonetheless, by highlighting potential hotspots of replication stress, our model provides a valuable starting point for deeper experimental investigations into the molecular basis of fragility. All tests and supporting code are provided in the GitHub repository: https://github.com/fberkemeier/DNA_replication_model.git.

## 2.5 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 analysed. 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. Supplementary Fig. 5 shows the correlations between replication timing error, firing rates, and transcriptional or chromatin features, demonstrating the relevance of these tests to our data.




**Supplementary Figure 5: 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 $< 10^{-15}$.

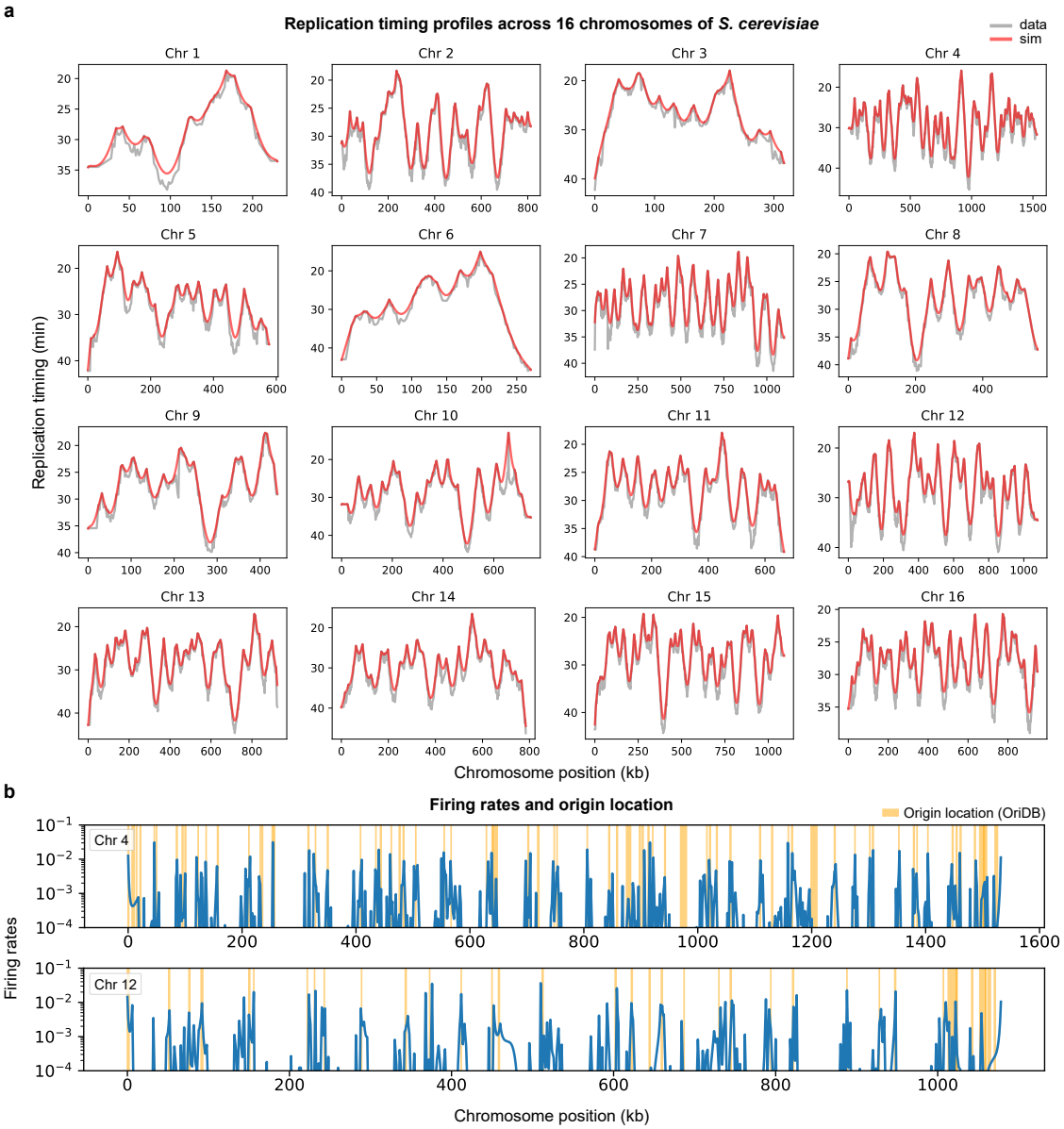
### 2.6 Theoretical digression: Application to *Saccharomyces cerevisiae*

Although our primary analyses focus on the human genome, the underlying framework is fully generalizable to other eukaryotes. To demonstrate this, we apply our fitting pipeline to *S. cerevisiae* (budding yeast) replication-timing data[16,17], where the most active origins are known[18]. In particular, we test whether our model is able to recover the well-established origins in yeast. To adapt to yeast's shorter genome, we constrain the neighbour-sum in Equation (1) for each site $j$ to

$$\max(1-j,-k) \leq i \leq \min(n-j,k) \tag{S22}$$

instead all $|i| \leq k$. Choosing the radius of influence $R$ to equal each chromosome's length then handles chromosome ends automatically, without altering any core assumptions.

To assess whether our model can recover known replication origins in yeast, we fit firing rates to replication timing data from Müller et al.[16] (Supplementary Fig. 6a), and compare the results to autonomously replicating sequences (ARS) annotations, independently obtained from the OriDB database[18], selecting those origins marked as 'Confirmed' or 'Likely'. This yields a one-dimensional profile of firing rates across the genome, along with a binary indicator vector specifying whether each genomic position falls within an annotated origin interval. We find that firing rates are systematically higher within these intervals (Supplementary Fig. 6b), and that the model recovers $> 86$ % of the origins at high-firing rates within $\pm$ 2 kb. To quantify this enrichment, we applied a Mann–Whitney $U$ test and a point-biserial correlation to evaluate the association between the binary origin label and the continuous firing rate. These tests produced highly significant results ($p \leq 10^{-12}$), indicating that the model successfully recovers regions of known origin activity.




**Supplementary Figure 6: Fitting the model in *S. cerevisiae*.**
**a** Observed replication timing from Müller et al.[16] compared with simulated timing profiles across the entire yeast genome, fitted using Equation (1) on each full chromosome. **b** Examples of fitted firing-rate profiles for chromosomes 4 and 12, highlighting sharp peaks at known origin locations from the OriDB database[18] ('Confirmed' and 'Likely'). The model infers these peaks from timing data alone, effectively suppressing firing activity in non-origin regions and recovering known origin locations without prior information.

# Supplementary Tables

**Supplementary Table 1:** Table listing the ten largest genes exhibiting misfits across all chromosomes, ranked from largest to smallest (left to right). Genes located at fragile sites are annotated as follows: C for common fragile sites, R for rare fragile sites, and CR for genes reported at both. All gene annotations refer to H1 cells with Repli-seq data aligned to the hg38 genome.

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

Note: Gene names are presented in italics following conventional nomenclature.

# References


1. Andrey N. Tyurin. Quantization, classical and quantum field theory and theta-functions. Preprint at https://arxiv.org/abs/math/0210466 (2002).

2. Nico M. Temme. *Error Functions, Dawson's and Fresnel Integrals* (NIST, Gaithersburg, 2010).

3. Michael A. Boemo, Luca Cardelli & Conrad A. Nieduszynski. The beacon calculus: a formal method for the flexible and concise modelling of biological systems. *PLOS Computational Biology* **16**, e1007651 (2020).

4. Devon Ryan *et al.* pyBigWig. Zenodo, https://doi.org/10.5281/zenodo.5144144 (2021).

5. R. Scott Hansen *et al.* Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences* **107**, 139–144 (2010).

6. Peiyao A. Zhao, Takayo Sasaki & David M. Gilbert. High-resolution Repli-seq defines the temporal choreography of initiation, elongation and termination of replication in mammalian cells. *Genome Biology* **21**, 1–20 (2020).

7. Lora Boteva *et al.* Common fragile sites are characterized by faulty condensin loading after replication stress. *Cell Reports* **32**, 108189 (2020).

8. S. G. Durkin, M. F. Arlt, N. G. Howlett & T. W. Glover. Depletion of CHK1, but not CHK2, induces chromosomal instability and breaks at common fragile sites. *Oncogene* **25**, 4381–4388 (2006).

9. Rajesh Kumar *et al.* HumCFS: a database of fragile sites in human chromosomes. *BMC Genomics* **19**, 985 (2019).

10. Grant R. Sutherland. Rare fragile sites. *Cytogenetic and Genome Research* **100**, 77–84 (2003).

11. J. K. Nancarrow *et al.* Implications of FRA16A structure for the mechanism of chromosomal fragile-site genesis. *Science* **264**, 1938–1941 (1994).

12. Seyed A. Hosseini *et al.* Common chromosome fragile sites in human and murine epithelial cells and FHIT/FRA3B loss-induced global genome instability. *Genes, Chromosomes and Cancer* **52**, 1017–1029 (2013).

13. Patrizia Vernole *et al.* Common fragile sites in colon cancer cell lines: role of mismatch repair, RAD51 and poly(ADP-ribose) polymerase-1. *Mutation Research / Fundamental and Molecular Mechanisms of Mutagenesis* **712**, 40–48 (2011).

14. Zaira M. Limongi, Angela Curatolo, Franca Pelliccia & Angela Rocchi. Biallelic deletion and loss-of-expression analysis of genes at FRA2G common fragile site in tumour-derived cell lines. *Cancer Genetics and Cytogenetics* **161**, 181–186 (2005).

15. Audesh Bhat, Parker L. Andersen, Zhoushuai Qin & Wei Xiao. REV3, the catalytic subunit of Pol $\zeta$, is required for maintaining fragile-site stability in human cells. *Nucleic Acids Research* **41**, 2328–2339 (2013).

16. Carolin A. Müller *et al.* The dynamics of genome replication using deep sequencing. *Nucleic Acids Research* **42**, e3 (2014).

17. Rosie Berners-Lee, Eamonn Gilmore, Francisco Berkemeier & Michael A. Boemo. Regulation of replication timing in *Saccharomyces cerevisiae*. Preprint at https://doi.org/10.1101/2024.10.000000 (2024).

18. Cheuk C. Siow, Sian R. Nieduszynska, Carolin A. Müller & Conrad A. Nieduszynski. OriDB, the DNA replication origin database updated and extended. *Nucleic Acids Research* **40**, D682–D686 (2012).