Published online 28 January 2021

NAR Genomics and Bioinformatics, 2021, Vol. 3, No. 1 **1**
doi: 10.1093/nargab/lqaa112

In silico analysis of DNA re-replication across a complete genome reveals cell-to-cell heterogeneity and genome plasticity

Maria Anna Rapsomaniki[1,2], Stella Maxouri[1], Patroula Nathanailidou[1], Manuel Ramirez Garrastacho[1], Nickolaos Nikiforos Giakoumakis[1], Stavros Taraviras[3], John Lygeros[2,*] and Zoi Lygerou[1,*]

[1]Department of Biology, School of Medicine, University of Patras, 26500 Rio Patras, Greece, [2]Automatic Control Laboratory, ETH Zurich, 8092 Zurich, Switzerland and [3]Department of Physiology, School of Medicine, University of Patras, 26500 Rio Patras, Greece

Received August 25, 2020; Revised December 15, 2020; Editorial Decision December 22, 2020; Accepted January 20, 2021

ABSTRACT

DNA replication is a complex and remarkably robust process: despite its inherent uncertainty, manifested through stochastic replication timing at a single-cell level, multiple control mechanisms ensure its accurate and timely completion across a population. Disruptions in these mechanisms lead to DNA re-replication, closely connected to genomic instability and oncogenesis. Here, we present a stochastic hybrid model of DNA re-replication that accurately portrays the interplay between discrete dynamics, continuous dynamics and uncertainty. Using experimental data on the fission yeast genome, model simulations show how different regions respond to re-replication and permit insight into the key mechanisms affecting re-replication dynamics. Simulated and experimental population-level profiles exhibit a good correlation along the genome, robust to model parameters, validating our approach. At a single-cell level, copy numbers of individual loci are affected by intrinsic properties of each locus, *in cis* effects from adjoining loci and *in trans* effects from distant loci. *In silico* analysis and single-cell imaging reveal that cell-to-cell heterogeneity is inherent in re-replication and can lead to genome plasticity and a plethora of genotypic variations.

INTRODUCTION

DNA replication ensures the maintenance of genetic information and constitutes the basis of biological inheritance. In eukaryotes, DNA replication initiates at multiple sites across the genome, known as origins of replication, and continues bidirectionally through replication forks that move continuously until precisely two DNA copies are produced (1,2). DNA replication is a complex and uncertain process, as only a small fraction of all putative origins is selected to fire in each cell, resulting in an individual progression along the genome at a single-cell level (3,4). Despite this high degree of stochasticity, DNA replication is also remarkably robust: it is tightly regulated in time and space by multiple control mechanisms that ensure its completion in an accurate and timely manner (1,5–7).

To maintain genome stability, each part of the genome must be replicated once and only once every time a cell divides. At the beginning of each cell cycle, two licensing factors, Cdt1 and Cdc6/18, load the MCM2–7 replicative helicase onto DNA, thereby licensing origins for a new round of DNA replication (8,9). In S phase, the replicative helicase either becomes active and moves away from origins with the replication fork or is removed by passive replication. Cdt1 and Cdc6/18 are strictly controlled and are inactivated as soon as replication starts, ensuring that the replicative helicase cannot load again onto origins that have been replicated, and therefore origins cannot fire a second time. Disruption of this control mechanism leads to re-firing of origins within the same cell cycle, a pathological process known as DNA re-replication (10). Overexpression of the licensing factors Cdt1 and Cdc6/18 has been shown to promote re-replication from yeast to humans. In fission yeast, overexpression of Cdc18 leads to origin re-licensing within the same cell cycle, origin re-firing and an uneven increase of DNA copy number (11), resulting in local amplification of the genome (12,13). Re-replication is enhanced by concomitant expression of Cdt1 (14,15). In budding yeast, the simul-

*To whom correspondence should be addressed. Tel: +30 2610 997689; Fax: +30 2610 991769; Email: lygerou@med.upatras.gr
Correspondence may also be addressed to John Lygeros. Tel: +41 44 632 8970; Fax: +41 44 632 12 11; Email: jlygeros@ethz.ch
Present address: Maria Anna Rapsomaniki, IBM Research Laboratory, Säumerstrasse 4, Rüschlikon 8803, Switzerland.

© The Author(s) 2021. Published by Oxford University Press on behalf of NAR Genomics and Bioinformatics.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

taneous inactivation of multiple control mechanisms leads to local re-replication (16,17), which can be converted to a local increase in gene copy number (18,19) and increased nucleotide-level mutagenesis (20). In mammalian cells, ectopic expression of Cdt1 alone or in combination with Cdc6 is sufficient to drive re-replication at multiple loci (21). Both Cdt1 and Cdc6 are often overexpressed in human tumors (22), and have been linked to genomic instability early in the tumorigenesis process (23), which drives oncogenesis (24–30).

Fission yeast has been used as a model organism for re-replication studies, as regulatable expression of a single factor (Cdc18) leads to genome-wide re-replication (11). Re-replication levels can be experimentally manipulated by regulating Cdc18 expression levels or co-expression of cofactors, and can range from a DNA content close to normal (2C, the DNA content of a normal G2 cell but resulting from uneven replication along the genome) to 32C (12–14). At the population level, re-replication in fission yeast progresses relatively evenly across the genome (12,13), while a small number of prominent loci are re-replicated above the genome mean. Common features of the central origins underlying these re-replicating 'hotspots' include AT-richness, early firing in a normal S phase and localization in large intergenic regions (13), features that also characterize efficient origins (31). Re-replication and normal S-phase origins largely overlap; however, notable differences between specific loci suggest that re-replication dynamics differ from normal replication.

To date, a number of mathematical and computational models of DNA replication in a number of organisms have been developed (32–41). However, the properties and underlying mechanisms of DNA re-replication across the genome remain unknown. Given the large number of origins along the genome and the stochasticity of origin firing (42–44), it is unclear how re-replication would progress along the genome in each individual cell in a re-replicating population and how local properties and genome-wide effects would shape its progression and the resulting increases in the number of copies of specific loci. While methods to analyze re-replicating DNA at the population level are available (12–13,16–17,45–46), analysis of re-replication dynamics at the single-cell level is currently lacking. Motivated by this gap in the literature, in this work we present a realistic, dynamic model of DNA re-replication exploiting stochastic hybrid systems. Stochastic hybrid systems combine discrete and continuous states and stochasticity (47) and have been successfully used to capture complex biological processes (39,48–50). Using as input experimentally determined origin measurements from fission yeast, the model allows the simulation of DNA re-replication genome-wide. Sensitivity analysis showed that the model is robust and consistent with experimental data genome-wide, allowing rules governing re-replication to be unveiled. *In silico* analysis combined with in-cell validation showed that re-replication profiles at the single-cell level are characterized by a high degree of heterogeneity. Re-replication can, with varying probability, occur anywhere in the genome and generate many diverse genotypes within a population.

MATERIALS AND METHODS

DNA re-replication model and simulations

A complete mathematical description of the model states, transitions and inputs is given in Supplementary Note 1. A graphical overview of the model states and transitions is given in Figure 1A-C, and model inputs and outputs are shown in Supplementary Figure S1. The model was implemented using MATLAB R2016b and the source code, generated data and extensively documented figure-generating scripts are available under an open-source license at: https://github.com/rapsoman/DNA_Rereplication. Monte Carlo simulations were executed on the HPC cluster of ETH Zurich.

Statistical methods and data analysis

Denoising of raw CGH data. For the denoising step, we experimented with various methods (moving average filter, linear polynomial filter, a quadratic polynomial filter and a Savitzky–Golay filter) and a variety of parameter values (e.g. span/window size, degree of the fitted polynomial). From all combinations, a quadratic polynomial fit with a span size of 80 units was chosen as the most appropriate, because of its ability to eliminate noise while fitting the shape and preserving the height of the peaks (Figure 2, top in black).

Peak finding. To locate the peaks, we implemented a simple peak finding method, which identifies as a peak all local maxima, i.e. all locations where the gradient of the signal changes sign. We also applied an intensity cutoff threshold and set it to 1 to eliminate the peaks whose intensity was below the genome mean. At the same time, a peak-matching step with a threshold of 40 kb was applied, so that local maxima in experimental and simulated data with a linear distance <40 kb were assigned to the same peak location. The algorithm identified 12 peaks on Chromosome I, 11 peaks on Chromosome II and 6 peaks on Chromosome III in the denoised experimental data, marked as dotted vertical lines in Figure 2 (Supplementary Table S2). These numbers were fairly robust with respect to the denoising method but depended largely on the span size and intensity of the cutoff value. The same process was followed for the simulated data and resulted in the identification of 11 peaks in Chromosome I, 5 peaks in Chromosome II and 6 in Chromosome III (Figure 2, Supplementary Table S2).

Analysis of peak overlap. To assess the overlap between peaks identified in experimental and simulated data, we repeated the same peak matching process using a null model; instead of the simulated peaks, we used 22 randomly picked genome coordinates and calculated how many coincide with the peaks found in the experimental data using the same window of 40 kb. The procedure was independently repeated 100 000 times; the median overlap score between the null model and experimental data across all repetitions was 2 out of 22 and the maximal overlap score was 10 out of 22, which occurred only once in all 100 000 repetitions. Very similar results were obtained when the above analysis was

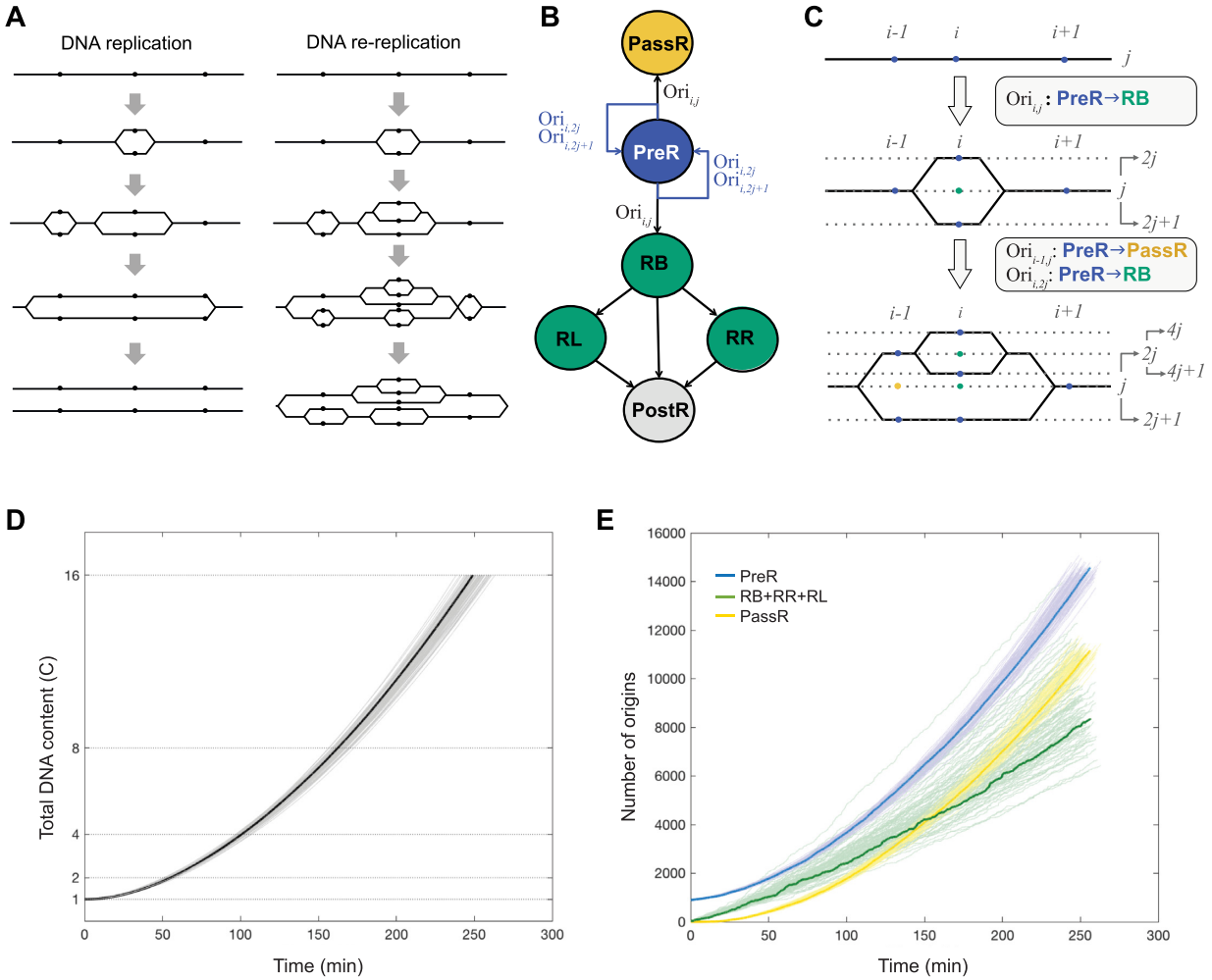

Figure 1. A stochastic hybrid model of DNA re-replication. (**A**) Normal DNA replication versus DNA re-replication. Normal DNA replication (left) starts from multiple replication origins (shown here as dots) and is tightly controlled, ensuring that during each cell cycle every origin fires once and precisely two DNA copies are produced. During DNA re-replication (right), re-firing of the origins results in many DNA copies on multiple strands and uneven amplification of the genome. (**B**) Abstract representation of the DNA re-replication model. Circles of different colors represent discrete origin states and arrows represent allowed transitions (black: transitions in normal replication; blue: transitions allowed only in re-replication). When an origin fires or is passively replicated, its offspring automatically fall into the $PreR$ state and can thus fire or be passively replicated again. (**C**) Evolution of re-replication and example transitions between states. Dots of different colors correspond to origins of different states (same as in B). Solid black lines represent synthesized DNA and dotted gray horizontal lines correspond to different strands (strand index shown in gray on the right). Initially, all origins pictured are in the $PreR$ state and located on strand $\boldsymbol{j}$. Then, $Ori_{i,j}$ fires and its offspring, $Ori_{i,2j}$ and $Ori_{i,2j+1}$ automatically fall into the $PreR$ state. Next, the left fork of $Ori_{i,j}$ reaches the location of $Ori_{i-1,j}$, which leads to its passive replication and the birth of origins $Ori_{i-1,2j}$ and $Ori_{i-1,2j+1}$ that automatically fall into the $PreR$ state. In the meantime, $Ori_{i,2j}$ also fires and creates $Ori_{i,4j}$ and $Ori_{i,4j+1}$ which again fall into the $PreR$ state. Note that the doubling of the strand index (starting with the original strand $\boldsymbol{j} = 1$) allows us to uniquely identify all strands. (**D**) DNA re-replication kinetics for 100 Monte Carlo simulations. Total DNA content (***C***) over time. Different curves correspond to different simulations. (**E**) Total number of origins per state over time. Different colors correspond to different origin states and different curves to different simulations. Highlighted curves correspond to the simulation closest to the mean.

repeated with 22 randomly selected origins (out of the 839 origins). Consistently, we calculate that the genomic regions assumed to be coinciding with an experimental peak have a total length of $29 \cdot 40\,001 = 1\,160\,029$ bases (29 experimental peaks with a surrounding 40 kb window each), corresponding to 0.0923 of the total genome (genome length = 12 571 820 bases). Randomly sampling 22 genome locations would lead to $22 \cdot 0.0923 = 2.0306 \approx 2$ re-replication peaks selected by chance alone, consistent with our numerical analysis of the null model above. The identified overlap between experimental and simulated data (14 out of 22 peaks) is therefore much higher than expected by chance.

Comparison of origin locations and efficiencies across datasets. A comparison between the origin locations and efficiencies used as input in our model (31) and an independent dataset by Daigaku *et al.* (51) was performed, to assess if disagreements between the two datasets could explain observed inconsistencies between experimental and predicted re-replication peaks. It should be noted that Daigaku *et al.* report efficiencies in the presence of passive replication, and not intrinsic efficiencies, and that reported efficiencies are overall higher in Daigaku *et al.* in comparison to (31). Nonetheless, comparing the location and relative efficiencies of assigned origins across problematic re-

gions in the two datasets can help pinpoint loci where incorrect origin assignment in our input dataset may be the cause of inconsistencies between simulated and experimental data. Out of the 15 experimental peaks missed by the model, 12 could potentially be attributed to a discrepancy in reported efficiencies; the model efficiencies used as input in the peak proximal region were relatively low, but the Daigaku *et al*. dataset includes at least one highly efficient proximal origin. Conversely, out of the eight peaks predicted by the model that are absent from experimental data, two could potentially be attributed to the opposite effect; the model input includes a highly efficient origin that is not present in the Daigaku *et al*. dataset. In Supplementary Table S3, a comparison across all experimental and simulated peaks together with the efficiencies of proximal origins in both datasets is presented. In Supplementary Figure S2, five representative examples are shown. Peak III-1 is a highly amplified region, corresponding to two overlapping re-replication hotspots from (13), which is not predicted by the model. The Daigaku *et al*. dataset reports two highly efficient origins, missing from our input dataset, which could explain the inconsistency. In contrast, in the same region our input dataset but not the Daigaku *et al*. dataset contains an efficient, left-telomere proximal origin, which underlies simulated peak SIII-1, missing from the experimental re-replication data. Similarly, for Peak SI-5, which is predicted by the model but not present in the experimental dataset, an underlying highly efficient origin from our input dataset was not validated by Daigaku *et al*. Dataset inconsistencies can also explain discrepancies between simulated and experimental re-replication data in the middle of Chromosome III, where the most pronounced simulated peak is less strong in the experimental data and vice-versa. Not all disagreements between experimental and simulated re-replication data can be explained by dataset discrepancies, however. For example, the underlying efficiencies for experimental Peak II-8, which is not predicted by the model, are low, whereas the underlying efficiencies for simulated Peak SII-3, which is not present in experimental data, are high for both datasets (Supplementary Figure S2).

Kernel density estimation. The distribution plots were derived using a kernel density estimate based on a normal kernel and evaluated at 100 equally spaced points.

Shannon entropy. To estimate the value of Shannon entropy H, we first discretized the data to 0 and 1, where 0 corresponds to copy numbers below the genome mean and 1 to simulations above the genome mean. Then, H is the entropy of a Bernoulli process with probability p of two possible and mutually exclusive outcomes, and is defined as follows:

$$H = -p\log p - (1-p)\log(1-p),$$

where $\Pr(X = 1) = p$and $\Pr(X = 0) = 1 - p$. Entropy will take its maximal value of 1 when $p = 0.5$, i.e. in the case that an origin is amplified in half of the simulations.

Principal component analysis. To compute the principal components of the data we used the MATLAB function implementation of Principal Component Analysis (PCA) and visualized the results (variable loadings and principal components) using a biplot.

Clustering. To identify groups of similarly amplified profiles in the simulated data, we performed a clustering step using the k-means algorithm with a squared Euclidean distance metric. We used the Gap statistic to identify the optimal k, a goodness-of-clustering approach that compares the change in within-cluster dispersion with that expected from a reference null distribution (52). We estimated the Gap statistic for up to $k = 50$ clusters using 100 reference data sets and selected as the optimal the smallest value of k for which the value of the Gap statistic is not more than 1 standard error away from the first local maximum. To estimate the stability of the clustering we used the Adjusted Rand Index (ARI), a pairwise metric of similarity between two different clustering assignments. The simple Rand Index (RI) is defined as the number of agreements over all pairs of samples between two different clustering assignments, divided by the total number of pairs, and reflects the probability that a randomly picked pair of samples is consistently found belonging to the same cluster. The ARI is an extension of the RI that additionally corrects for chance (53). To compute the ARI, we ran 100 repetitions of the clustering for the optimal k identified as described above, where each repetition was an independent run with a random centroid initialization.

Implementation. All methods were implemented in the Statistics and Signal Processing Toolboxes of MATLAB 2016b.

Experimental methods

Strain construction and cell growth. For the construction of the re-replicating strain with the *lys1+* tag the strain C2566 (h-, LacO::lys1+, LacI-GFP::his7+, ChrI 1.5Mb::TetO-hphMX, Z locus::TetR-tdTomato-natMX, leu1–32, ura4-D18, ade6-M210) kindly provided by Christian Häring was employed (54). To promote re-replication C2566 was crossed with a strain carrying a truncated form of *cdc18* under the nmt1 promoter (*h-, nmt1-d55P6 (leu1+), ura4-D18, leu1–32*) (55). The resulting strain undergoes medium level of re-replication, which is adequate for the manifestation of high cellular heterogeneity but does not severely decrease the viability of the cells. Moreover, medium level of re-replication does not affect the integrity of the nucleus, thus it is possible to detect and quantify Hoechst staining through imaging. This strain was used for the quantification of the *lys1+* locus. To induce re-replication, cells were grown at 25°C in Edinburgh Minimal Medium (EMM) supplemented with uracil and adenine in the presence of thiamine (5 μg/ml) up to an OD600∼0.5. At this point cells were harvested, washed with EMM and diluted to an OD600∼0.01 in EMM supplemented with uracil and adenine, both with and without thiamine and grown at 25°C for 30 h. Cells were harvested after 30 h, fixed with 4% paraformaldehyde for 5 min, washed with water and stained with Hoechst for 5 min.

Cell imaging and image analysis. Cell images were obtained in an Olympus IX83 widefield microscope equipped with a 100× lens (NA 1.46) and a LED light-source. Z-stacks were acquired at a step size of 0.4 μm. Preprocessing, segmentation and signal quantification for GFP and Hoechst were conducted with ImageJ. The top-hat algorithm was applied for background correction. Signals for each channel individually were detected by manual thresholding.

RESULTS

Modeling DNA re-replication across a complete genome

DNA replication initiates from hundreds of origins along the genome and results in the exact duplication of the genetic material (Figure 1A). It is a complex process that involves a combination of discrete dynamics (associated with the switch-like activation of each origin), continuous dynamics (associated with the movement of the replication forks along the DNA strands) and stochasticity (in the time and space of origin firing). How often a putative origin is observed to fire in a population of normally replicating cells depends on how often it is licensed for replication, how often it is activated to fire when licensed, and how often it is passively replicated. Here, we define as intrinsic firing efficiency the fraction of cells where an origin would be observed to fire in the absence of passive replication. Intrinsic firing efficiency thus encompasses both the ability of the origin to become licensed and its ability to fire. We define as firing propensity the probability that an origin fires in a unit of time. Mathematical and computational models of DNA replication have been proposed in the literature to capture origin firing in time and space and the dynamics of DNA replication (32,34–35,37,39,41,56). We developed an extension of a mathematical model of normal DNA replication (39) to allow origin re-firing, resulting in a stochastic hybrid model of DNA re-replication.

Contrary to normal DNA replication, where exactly two DNA copies are produced, in re-replication origin re-firing allows each origin to produce multiple copies (referred to here as offspring) on multiple resulting strands (Figure 1A). Each origin copy, whether ancestral or offspring, can be identified by its genomic location and strand index and is denoted as $\mathrm{Ori}_{i,j}$, where $i = 1, \ldots, n$ denotes the origin index and $j = 1, \ldots, m$ the strand index. Figure 1B pictorially summarizes the discrete dynamics of the model. Similarly to the model of (39), at any point in time, each origin can be in one of the depicted six states: pre-replicative ($PreR$), replicating in both directions (RB), replicating only to the right or to the left (RR or RL), passively replicated ($PassR$) and post-replicative ($PostR$). The continuous dynamics are deterministic and model the movement of the replication forks. Uncertainty plays a vital role in re-replication, and it is represented by modeling the time and location of origin firing and re-firing as stochastic events. Transitions between discrete states, depicted as arrows in Figure 1B, depend on both the continuous and stochastic dynamics of the system. In contrast to normal DNA replication, origins that have already fired or have been passively replicated can re-fire multiple times. These transitions are visualized in an example scenario in Figure 1C, when an origin fires (transition $PreR \rightarrow RB$) or is passively replicated (transition $PreR \rightarrow PassR$), it generates two origins on two new strands, which automatically fall back into the $PreR$ state and can thus fire again (blue arrows in Figure 1B). To assign firing propensities of these newly replicated origins, we assume that the total firing propensity (the sum of all firing propensities of every origin in the cell) remains constant (see limiting factor hypothesis (39) and below for alternative implementations). Each time an origin fires (or is passively replicated), its firing propensity gets dynamically redistributed to all pre-replicative origins across all strands, in proportion to their current firing propensity.

The DNA re-replication model requires the following inputs (Supplementary Figure S1): (i) total genome length, measured in base pairs, (ii) genomic locations of all origins, measured in base pairs, (iii) intrinsic firing efficiencies of all origins and (iv) fork speed, measured in kilobases replicated per minute (kb/min). Provided that this information is available, the model is applicable to any eukaryotic genome. For the purposes of this work, the model was instantiated for the case of fission yeast (*Schizosaccharomyces pombe*). This organism has long served as a model system for the study of DNA replication control, as it exhibits conserved features, while its small genome (total genome length $\approx 14 \times 10^6$ bases and three chromosomes) simplifies analysis. Exact origin locations and their intrinsic firing efficiencies (fraction of origins that fire when fork movement is blocked with hydroxyurea) have been measured experimentally across the complete fission yeast genome during normal replication (31) by microarray analysis. Supplementary Table S1 shows the locations and efficiencies of the 893 fission yeast origins used as input. Each origin (Ori) is named for the chromosome on which it resides (I-III) and a sequential number along the chromosome.

We have assumed a constant fork speed. Though the model can accommodate different fork speeds at different chromosomal locations or across time, experimental data for assigning a variable fork speed are currently lacking. While initial estimations of mean fork speed during normal DNA replication were around 3 kb/min (31,57–58), recent estimations of fork speed range from 0.5 to 1.5 kb/min (59,60). Mean fork speed in re-replication is expected to be slower than normal replication, due to limited nucleotide pools and interference between forks. Therefore, in our base-case model, we set fork speed equal to 0.5 kb/min and assumed it is uniform across the genome. In contrast to normal DNA replication, where the process is completed when all genomic regions have doubled, in re-replication there is no defined endpoint. The ploidy level C, i.e. the total amount of genomic material synthesized with respect to the initial amount, can be used to define discrete points along the process. It should be noted that re-replication is not expected to progress evenly across the genome and therefore different genomic regions will be amplified to different extents at a given time point.

Monte Carlo simulations of the model for the aforementioned inputs were used to study the re-replication process. Since the model is stochastic, each simulation corresponds to a sample path of the stochastic process, i.e. a distinct sequence of random events; Monte Carlo simulations permit the estimation of statistics over multiple such sequences.

Due to the complexity of the process, the discrete and continuous state space quickly becomes very large. In a normal DNA replication cycle, the total number of origins doubles and reaches 1786 as the total DNA content C increases from 1 to 2, while the maximum number of active forks will be up to double the number of origins. During the course of re-replication, however, the number of origins and forks, equivalent to the discrete and continuous states of the model, increases drastically, in an example simulation, already by the time that C reaches 2, 3861 origins at various states and 2515 active forks are present. For $C = 16$, the count of origins and forks escalates and increases almost 10-fold, with 35 259 origins and 18 827 active forks present.

In Figure 1D, examples of the kinetics of DNA synthesis over time are shown. Each curve corresponds to a single simulation, and the spread between individual curves indicates variability due to the stochastic nature of the model. We observe that the increase in DNA content over time is exponential, and a DNA content of $8C$ is reached within ∼3 h, in the same range as experimental observations (12,13). The number of active (RB, RR, RL), passive ($Pass\,R$) and pre-replicative origins ($Pre\,R$) over time are shown in Figure 1E. During the course of re-replication, the number of active and passive origins increases rapidly over time. Initially, passively replicated origins are fewer than the actively replicated ones, but as re-replication progresses, the number of passively replicated origins increases faster and surpasses the number of origins that fired. This indicates that the process gets eventually dominated by passive replication instead of firing events. Pre-replicative origins increase exponentially, as all firing and passive replication events lead to the birth of new $Pre\,R$ origins.

We have therefore developed a model which can capture re-replication dynamics across an entire genome, accounting for transitions in origins states, fork movement and stochasticity.

DNA re-replication at a population level

Simulation results from the re-replication model were compared to experimental data from re-replicating fission yeast cells. Specifically, we computed *in silico* mean amplification profiles across the genome, referred to as signal ratios in (13), by averaging the number of copies for each origin location and normalizing it to the genome mean in 100 simulations. In these profiles, peaks above 1 correspond to highly re-replicated regions, and valleys below 1 correspond to regions that are under-replicated with respect to the mean. Mean profiles computed at $16C$ are shown in Figure 2, bottom row. Note that re-replication levels are higher on Chromosome III in comparison to the other two chromosomes, consistent with a higher efficiency of origins on this chromosome (31). Simulated re-replication profiles were compared to re-replication profiles defined experimentally at a similar ploidy (13), where location-specific amplification was assessed using array Comparative Genomic Hybridization in fission yeast cells co-overexpressing the licensing factors Cdc18 and Cdt1 (Figure 2, top row). Simulated data show the actual number of copies generated and are thus expected to be sharper than experimental data, which are subject to background noise and represent averages of three probes and two independent experiments. Still, *in silico* and experimental profiles appear overall similar, with several peaks coinciding. Indeed, our model predictions fit experimental observations reasonably well, as the Spearman correlation coefficient ρ between experimental and simulated whole-genome re-replication profiles was statistically significant for all three fission yeast chromosomes ($\rho = 0.6$ and P-value $= 3.6 \cdot 10^{-41}$ for Chromosome I, $\rho = 0.61$ and P-value $= 5.7 \cdot 10^{-33}$ for Chromosome II, and $\rho = 0.5$ and P-value $= 7.3 \cdot 10^{-12}$ for Chromosome III). To better compare simulated and experimental profiles, a peak-calling algorithm was used, which identified 29 and 22 peak locations in experimental and simulated profiles respectively (dotted vertical lines), representing regions of amplification in a population of re-replicating cells. Details on the denoising and peak finding processes are given in 'Materials and Methods' section, and indices and locations of both peak sets are given in Supplementary Table S2.

We observed that most peaks predicted from simulations correspond to major amplification peaks in the experimental data; out of the 22 re-replication peaks in the simulated data, 14 also exist in the experimental data (precision ≈ 0.64). To further assess the agreement between experimental and simulated peaks and examine whether it could be attributed to chance or our peak-matching algorithm parameters, we repeated the analysis 100 000 times using a null model of 22 randomly picked genome locations ('Materials and Methods' section). We found that there was no case out of the 100 000 random repetitions, with 14 peaks overlapping in experimental and randomly picked locations, indicating that the probability of the model's prediction being attributed to chance is less than 1 in 100 000. The maximal overlap, occurring only once in all 100 000 random repetitions, was 10 peaks, while the median overlap between experimental and random peaks was 2 peaks. Moreover, out of the nine amplification regions identified in (13), six are also present in the simulated profiles.

Some inconsistencies do exist between the two datasets; out of 29 peaks in the experimental data, 15 are not predicted (false negative rate ≈ 0.52). Most of these false negatives however are attributed to peaks that are present but appear less sharp in the simulations (e.g. peaks in the middle part of Chromosome II) or to minor disagreements due to linear shifts in peak locations (e.g. the first and third true peaks of Chromosome III (31)). To assess if some of these discrepancies could be due to incorrect assignment of origin locations or efficiencies in the input dataset, we compared the input data with a different origin dataset, estimated using a polymerase usage sequencing (Pu-seq) strategy during an unperturbed cell cycle in *S. pombe* (51). In this dataset, passive replication is not inhibited, and efficiencies therefore do not correspond to intrinsic firing efficiencies, as required for model input. It can nevertheless be used to pinpoint regions where inconsistencies could be explained by the input data. Indeed, minor disagreements across datasets could account for observed inconsistencies in several cases (Supplementary Table S3 and Figure S2; 'Materials and Methods' section). For example, the linear shift of the first peak on chromosome III in experimental and simulated data mentioned above can be explained by such a disagreement in

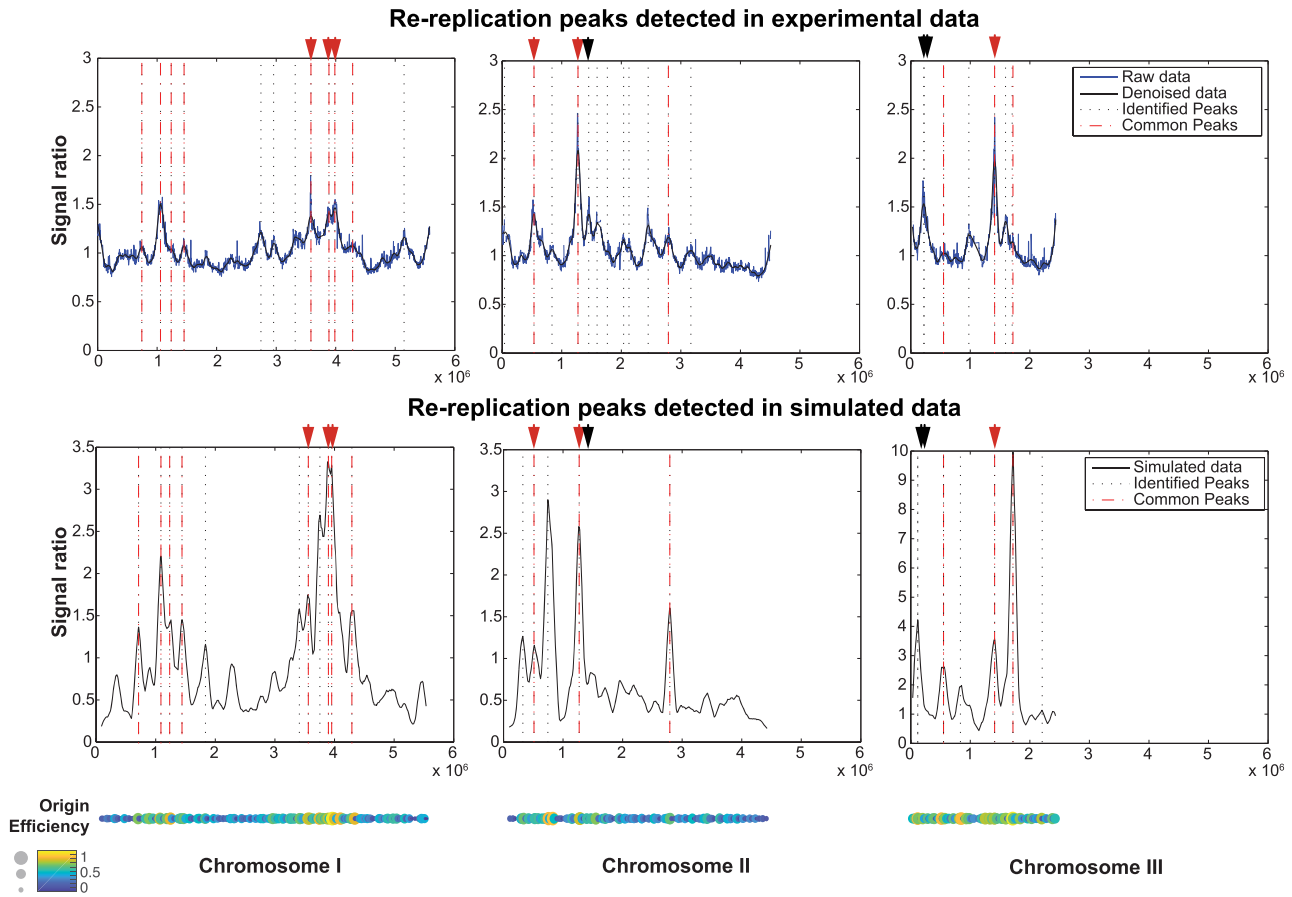

Figure 2. Analysis of *in silico* data at a population level. Comparison between experimental (top row) and simulated (bottom row) mean amplification profiles for 100 Monte Carlo runs for 16***C*** along the three chromosomes of the fission yeast genome. Identified peaks, representing re-replication hotspots, are marked in dotted vertical lines. Common peaks between simulated and experimental data are marked in red dashed vertical lines. Peaks corresponding to the nine amplification regions identified by Kiang *et al*. (13) are marked in vertical arrows (red: identified also in simulated data, black: identified only in experimental data).

efficiency datasets (Supplementary Figure S2). Striking inconsistencies which cannot be accounted for by discrepancies in the input dataset are relatively few and can point to regions along the genome which are specifically regulated under re-replication conditions. For example, the third peak on chromosome II in the simulated data (Peak SII-3) is supported by both efficiency datasets (Supplementary Figure S2) but is not detected in experimental re-replication data, suggesting that re-replication of this locus may be inhibited in cells. Similarly, the eighth peak on chromosome II (Peak II-8) identified in experimental data and absent in simulated data resides in a low efficiency region in both origin datasets available (Supplementary Figure S2) and could indicate a region particularly prone to re-replication. Notably, subtelomeres are also highly amplified in experimental data but not in the simulations, pointing to location-specific effects, not explicitly specified in our model, as previously suggested (12). Such isolated events however do not significantly affect the overall re-replication dynamics.

We conclude that simulated population data fit experimental data genome-wide reasonably well, validating our approach.

Sensitivity analysis

We next sought to investigate the effects of different model parameters and assumptions. We first varied fork speed: 0.5 kb/min (base case) was compared to 1 and 3 kb/min (Supplementary Figure S3). As expected, the increase in DNA content progresses faster at higher fork speeds: $C = 16$ is reached at 251.5, 187 and 114 min as fork speed increases from 0.5 to 1 and 3 kb/min (Supplementary Figure S3A and B). Experimental observations show that under high levels of re-replication, DNA content reaches 16*C* 4–6 h following accumulation of Cdc18 (14). Estimates for 0.5 kb/min are therefore closer to experimental observations. Passive replication becomes more dominant as fork speed increases (Supplementary Figure S3C).

A second variant of the model was tested, that differs on how the firing propensities of newly replicated origins are assigned. In this variant (referred to as Unlimited Factor or UF) we assume that when an origin fires or is passively replicated, the offspring inherit the same firing propensity as the parent. This implies that firing propensities depend only on the genomic location and hence remain the same for the same origin across all strands. Under this assumption the

total firing propensity (i.e. the sum of the propensities of all the origins) will increase during re-replication (Supplementary Figure S3D). By contrast, in the base-case (referred to hereafter as Limiting Factor or LF), the total system firing propensity remains constant during re-replication while the firing propensities of individual origins decrease, as more origins are born (Supplementary Figure S3E). The two variants of the model reflect different biological hypotheses. The UF variant represents a situation where all factors needed to license and activate an origin are available in virtually unlimited quantities. The LF variant represents a situation where one or more of these factors exists in limited quantities and binds to origins proportionally to their intrinsic efficiencies (39,61).

Simulation kinetics for the UF variant at a fork speed of 3 kb/min are shown in Supplementary Figure S3E. When unlimited copies of an activation factor (UF variant) are assumed, the process is fast, as DNA content doubles approximately every 12 min and reaches 16*C* in <1 h. For the LF variant, on the other hand, each doubling needs gradually more time to complete, and re-replication requires roughly twice as much time to reach the same *C* levels as in the UF case. In the UF model, the number of passive and active origins increases rapidly with a comparable count (Supplementary Figure S3C), in contrast to the LF model where active origins increase at a slower rate and are eventually outnumbered by the passive ones. The re-replication process predicted by the UF model is much faster than experimentally observed (Supplementary Figure S3B), suggesting that unlimited re-replication is unlikely to take place within cells.

Mean amplification profiles across the genome for different values of fork speed and the LF model variant are shown in Supplementary Figure S4 at 16*C*. Profiles appear flatter as fork speed increases, consistent with increased passive replication. The sites of over-amplification however appear at similar locations along the genome, suggesting that re-replication dynamics along the genome are robust to varying fork speeds. In Supplementary Figure S4, amplification profiles genome-wide are also compared between the base-case model (LF variant, fork speed of 0.5 kb/min) and the UF variant (fork speed of 3 kb/min). We observe that both profiles follow a very similar pattern, with the UF profiles characterized by somewhat sharper peaks, indicating more firing from the underlying origins. Importantly, all amplification peaks shown in Supplementary Figure S4 are consistently present in both model variants and parameter values.

We conclude that the base-case model is robust to model assumptions.

Single-cell analysis: heterogeneity across the genome

Re-replication across the genome has only been studied so far at the population level. Although population-based methods enable the exploration of global characteristics, they mask the underlying variability at a single-cell level, as only the most prominent regions 'survive' in the mean amplification profiles. The model described here permits analysis of cell-to-cell heterogeneity of the amplification levels across the genome, as each simulation corresponds to a distinct sequence of events taking place within one cell.

To assess variability at the single-cell level, we compared amplification plots from single simulations, generated by the base-case model. To assess the outcome at different ploidy levels, we compared simulations at 2*C* and 16*C*. In both cases, single-cell profiles are characterized by a high degree of variability and can deviate significantly from the mean behavior (examples of four random simulations at 16*C* in Supplementary Figure S5). To quantify the variability in the simulations genome-wide, we used the Shannon entropy, an information-theoretic metric ('Materials and Methods' section) on discretized data from 100 simulations, where 0 and 1 correspond to copy number levels less and more than the genome mean, respectively. As shown in Supplementary Figure S6, this analysis indicates that whether an origin is amplified or not is highly unpredictable at 2*C*, while at 16*C* the entropy becomes bimodal, with half of the origins consistently over- or under-replicated and the other half showing a highly variable behavior.

In Figure 3, a zoom in on 1 Mb of Chromosome I is shown at 2*C* (Figure 3A) and at 16*C* (Figure 3B) for the same four randomly selected individual simulations as in Supplementary Figure S5. We observe that amplification levels along the genome vary across the simulations, pointing to a high degree of heterogeneity. This variability is especially prominent early on in the process (2*C*), where certain origins have been amplified to a high degree, while most of the genome remains normal. We analyzed the copy number levels at 16*C* of an individual origin in this region (red dot in Figure 3A and B), for which a high degree of heterogeneity was not apparent in the particular simulations selected. We observed that, when looking at all simulations, the distribution (Figure 3C) is right-skewed with a heavy tail, showing that high variability in copy number levels is indeed present.

To determine whether every region along the genome is amenable to amplification upon re-replication, we computed the number of times each origin was amplified more than the genome mean (16*C*). This analysis showed that 739 out of the total 893 origins are amplified above the mean at least once in the set of 100 simulations analyzed. This suggests that even regions of low efficiency can potentially be amplified. A striking example is given in Figure 3D and E for origins Ori II-132 and Ori II-153. As seen from their copy number distributions (Figure 3D) and the mean re-replication profile (Figure 3E), both origins are underrepresented in the population and reside in a region of almost no re-replication. However, as seen in the single-cell profile, they can potentially re-replicate, and yield copies high above the population mean. This suggests that under re-replication, multiple combinations of co-amplified regions will appear, even for low efficiency regions.

We conclude that re-replication can drive different genomic regions to be amplified in different cells, leading to heterogeneity at the single cell level.

Single-cell heterogeneity observed *in vivo*

To experimentally explore the cell-to-cell copy number variability under re-replication *in vivo*, the relative amplification level at a specific genomic region was assessed using the LacO/LacI system. Specifically, high affinity binding of an ectopically expressed, fluorescent-tagged lactose

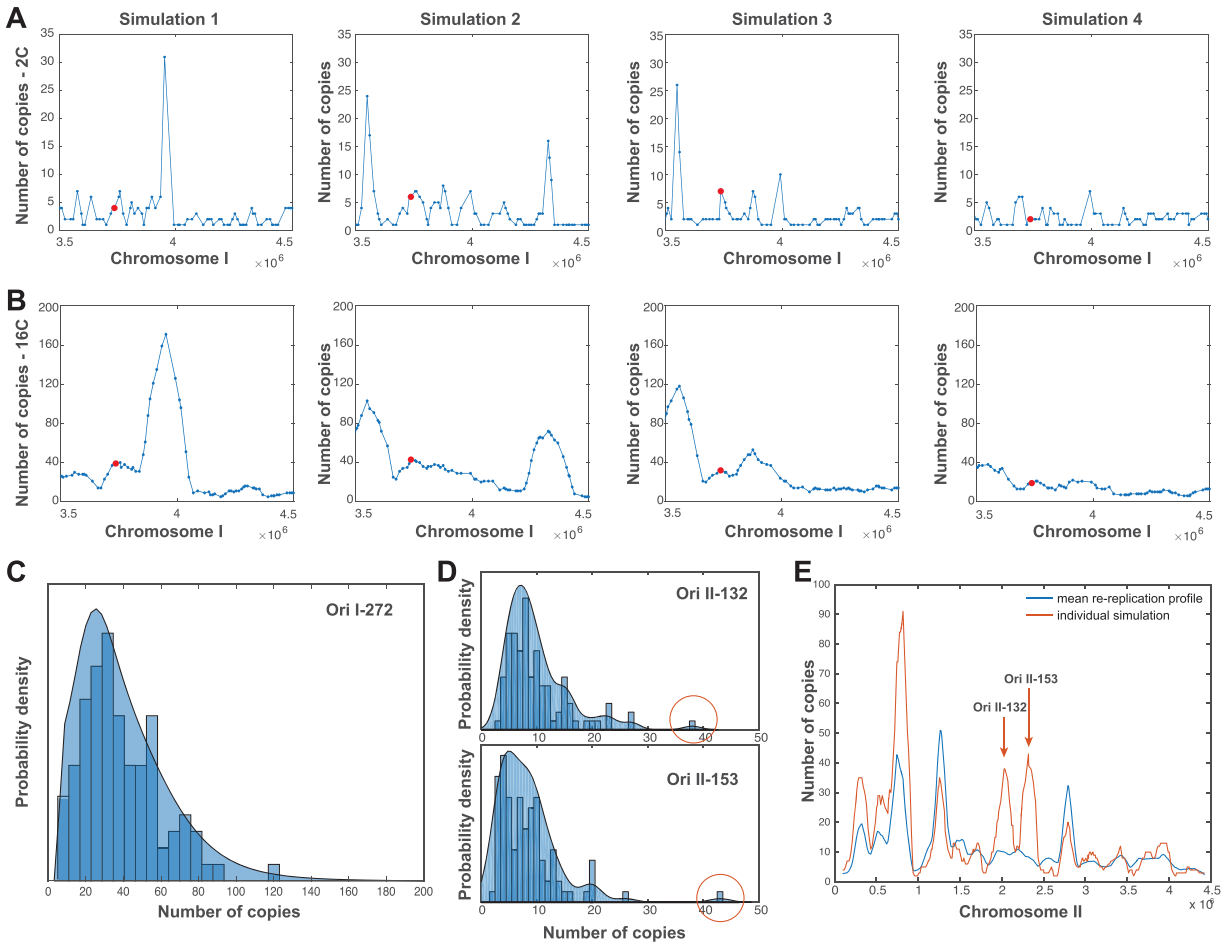

Figure 3. Analysis of *in silico* data at a single-cell level exposes heterogeneity. (**A** and **B**) Model simulations expose heterogeneous patterns of re-replication at a single-cell level. Individual simulations of the stochastic model lead to markedly different amplification levels. Shown here are number of copies for all origins (marked in circles) on a random region of Chromosome I, resulting from four random simulations of the model at a total DNA level of 2***C*** (A) and 16***C*** (B). (**C**) Copy number distribution of one individual origin of Chromosome I (Ori I-272), highlighted in red in (A and B), from the 100 simulations at 16***C***. (**D**) Distributions of copy number levels of weak origins Ori II-132 and Ori II-153. Outliers of the distributions are marked in circles. (**E**) Mean amplification profile of Chromosome II from all 100 simulations (blue) versus single-cell amplification profile corresponding to one individual simulation for which Ori II-132 and Ori II-153 are amplified.

inhibitor (lacI-GFP) onto stably integrated lac operator (lacO) arrays allows the visualization of a targeted genomic region as a fluorescent dot, the intensity of which reflects the copy number of the lacO-targeted region (62). To induce re-replication in a controllable manner, a fission yeast cell strain stably expressing the licensing factor Cdc18 under the repressible promoter nmt1 was employed. Absence of the vitamin B1, thiamine, activates the promoter and leads to Cdc18 overexpression. Different promoter constructs and different Cdc18 mutants have been described which can induce re-replication to varying degrees (from a 2*C* to 32*C* DNA content (14,63)). To avoid artifacts due to cell death and disrupted nuclear morphology under high levels of re-replication, we have employed a truncated form of Cdc18 (*d55P6-cdc18,* (55)) which induces medium-level re-replication, as confirmed by flow cytometry analysis in Supplementary Figure S7A. Under these conditions, the vast majority of the cells (>90%) undergo re-replication, albeit at medium to low levels (Supplementary Figure S7A and data not shown). Additionally, the same strain carries the *lys1+* locus marked by the lacO-lacI system (Figure 4A) (54). The *lys1+* gene is located between Ori I-272 and Ori I-273, which present 60 and 39% efficiency, respectively. Copy number levels for Ori I-272 in individual simulations and across the whole population were shown in Figure 3A–C.

Re-replication was induced by removing thiamine for 30 h at 25°C, or not as a control, and the cells were fixed, stained with the DNA dye Hoechst and imaged in a wide-field epifluorescence microscope (Figure 4B). Image analysis revealed an increase in DNA nuclear staining in the re-replicating cells (Figure 4C), as well as an increase in the intensity of the lacI-GFP dot (Figure 4D), each of them indicating increased genomic content and increased copies of the *lys1+* locus under re-replication, respectively. As shown in Supplementary Figure S7B, in control cells the distribution of DNA nuclear staining is consistent with the presence of G1, S and G2 phase cells, with G2 cells having approximately double the DNA content of G1 cells, while lacI-GFP foci intensities correlate with the DNA content. On the contrary, re-replicating cells do not present a clear sep-

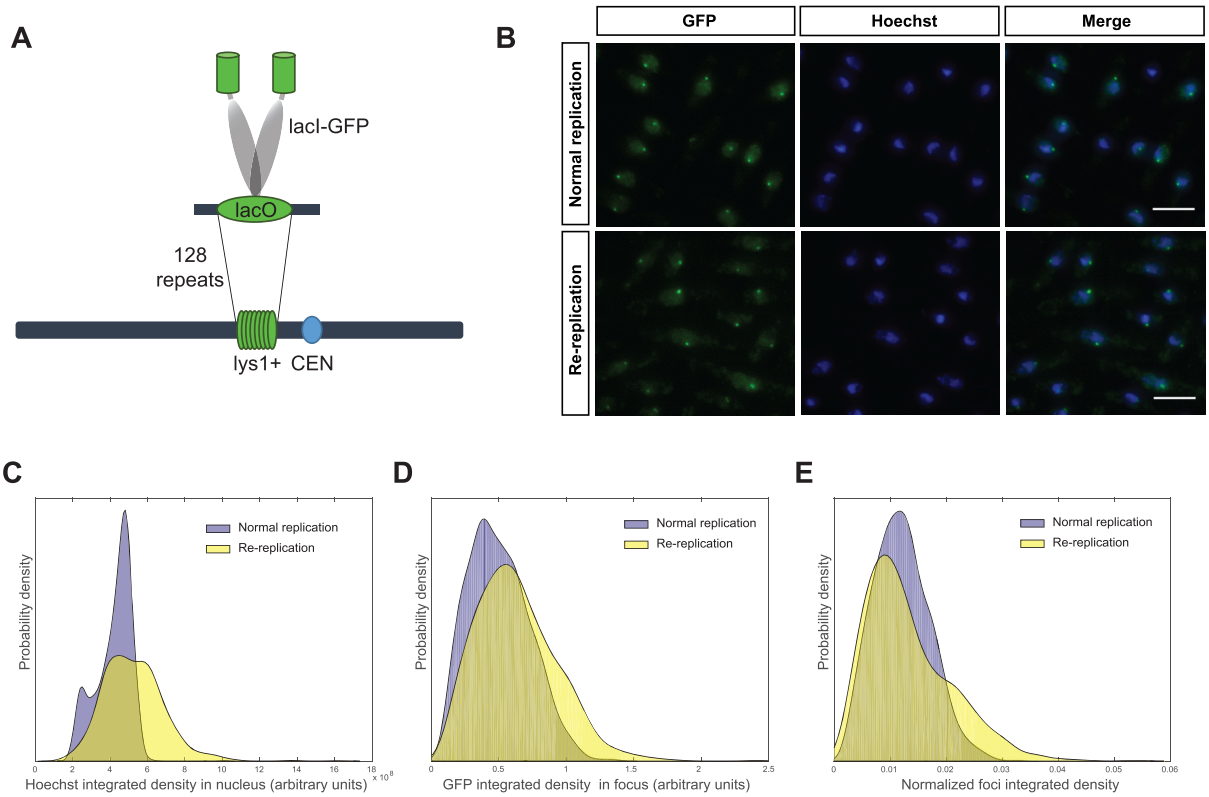

Figure 4. Quantification of lys1+ region in normal and re-replicating conditions. (**A**) A region proximal to the lys1+ gene (ChrI: 373492) was labeled by integration of 128 LacO repeats in fission yeast cells expressing LacI-GFP. (**B**) Visualization of the GFP labeled region in cells growing under normal replication and re-replication conditions. Cells were grown for 30 h at 25°C in the absence of thiamine to induce overexpression of d55P6-cdc18 and re-replication (lower panels) or were grown in the presence of thiamine as a control (upper panels). GFP (green) and DNA stained by Hoechst (blue) were visualized by epifluorescence microscopy. Scale bar: 5 μm. (**C–E**) Distributions (probability density plots) of total nuclear Hoechst intensity (C), GFP intensity at the lacO locus (D) and GFP focus intensity normalized to the total nuclear Hoechst intensity in each cell (E) is shown in cells undergoing normal replication ($n = 1632$, blue) or re-replication ($n = 1234$, orange), as in B. A representative experiment out of three biological replicates is shown.

aration of populations and varying levels of re-replication are observed in different cells, consistent with flow cytometry data (Supplementary Figure S7A). Foci intensities appear to vary independently of the DNA content. To estimate the relative copy number of the *lys1+* region with respect to the DNA content at the single cell level, the intensity of each GFP dot was normalized with the total DNA nuclear intensity in each individual cell (Figure 4E). We observe that under re-replication the distribution of the normalized GFP intensity is positively skewed with a long tail and an increased coefficient of variation compared to the normal replicating sample (41.18% for normal and 57.54% for re-replication), in agreement with the simulated data at this region (Figure 3C). We conclude that cell-to-cell heterogeneity in the number of copies of the *lys1+* genomic locus is evident in fission yeast cells undergoing re-replication, consistent with *in silico* analysis.

Rules governing DNA re-replication across the genome

Intrinsic origin properties. To unveil the rules which dictate which regions will become amplified along the genome, we first assessed the dependence of amplification levels of individual origins on their intrinsic efficiencies, as determined experimentally (31). Figure 5A shows histograms of the copy number distributions from simulations of the LF model at 0.5 kb/min at 16*C* of two origins (Ori II-45 and Ori II-54), with high and low efficiencies (62 and 9%, respectively (31)). The median number of copies of each origin is consistent with its efficiency (notice again the positively skewed distribution with long tails discussed above). The scatterplot of Figure 5B shows a strong correlation between mean number of fires at 16*C* and efficiency for all origins (Spearman correlation coefficient $-\rho = 0.96$). The coefficient of variation (ratio of standard deviation over the mean) is inversely correlated to the efficiency ($\rho = -0.89$), with weak origins showing much higher variation than strong ones. Mean number of copies of each origin show a weaker correlation to firing efficiency (Figure 5C, $\rho = 0.4$) and a coefficient of variation weakly linked with efficiency ($\rho = 0.12$). Interestingly, a higher spread at low efficiencies is observed, with origins considered dormant (efficiencies <10%) occasionally significantly amplified with respect to the population mean. Last, a scatterplot of the mean number of passive replications versus the efficiency (Figure 5D) indicates a much weaker correlation ($\rho = 0.17$) and a weak efficiency-related variability ($\rho = 0.13$). This analysis shows that re-firing of a given origin is strongly af-

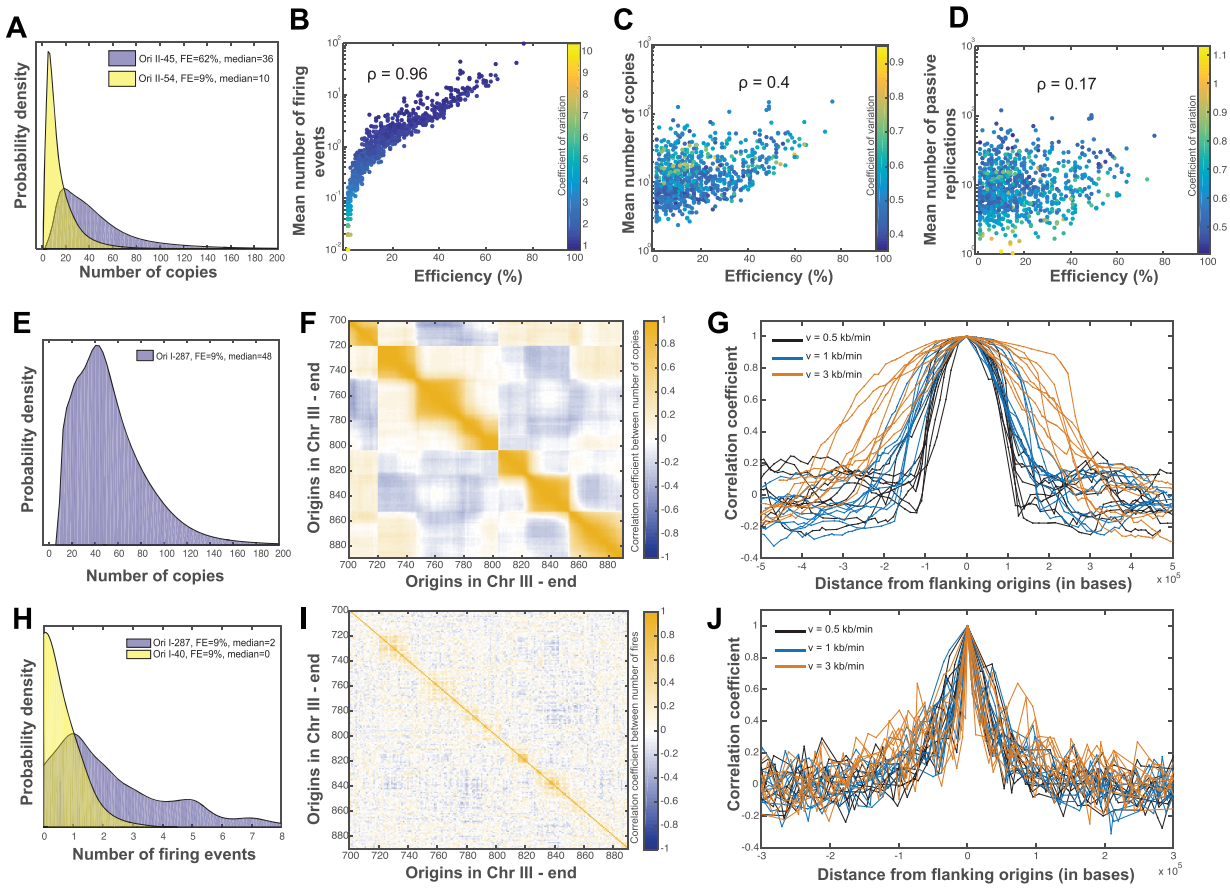

Figure 5. Analysis of *in silico* data at a single-cell level linked with intrinsic properties and points to *in cis* effects. Simulation results from 100 Monte Carlo simulations of the LF model at 0.5 kb/min at 16*C*. (**A**–**D**) Amplification levels of individual loci with respect to their intrinsic properties. (A) Distributions of copy number levels for two origins of high (purple) and low (yellow) efficiency (origin indices, efficiencies and median number of copies given in the legend). (B) Scatterplot of mean number of fires versus firing efficiency for all origins shows a strong correlation between firing events and firing efficiency (Spearman correlation coefficient value $-\rho = 0.96$). Color indicates coefficient of variation (standard deviation/mean), and points to higher variability in firing for the weak origins. (**C**) Scatterplot of number of copies of individual origins versus their firing efficiencies shows a weaker correlation ($\rho = 0.4$) and variation less dependent of efficiency. (**D**) Scatterplot of mean number of passive replications versus firing efficiency shows a very low correlation ($\rho = 0.17$). (**E**–**G**) Amplification levels of individual loci with respect to local effects. (E) Distribution of copy number levels of weak origin I-287, residing next to strong origin I-288, shows elevated levels due to passive replication. (F) Heatmap of correlations of copy number levels between different origins exposes strong *in cis* effects, as shown here for a zoomed in region in the end of Chromosome III. Color indicates Spearman correlation coefficients. G Correlation coefficients between copy number levels of the 10 origins with the highest intensity (different lines) and their neighboring origins, centered and zoomed in the origin locations. Different color indicates varying values of the fork speed. (**H**–**J**) Firing activity of individual origins with respect to local effects. (H) Distributions of firing events for two origins of low efficiency, with efficient (purple) and inefficient (yellow) neighbors, shows that weak origins fire more often when residing next to strong ones. (I) Same as in (F) but showing correlations between number of fires across the genome. (J) Same as in (G) but showing correlations between number of fires of prominent origins and their neighbors.

fected by its efficiency, while additional properties govern levels of amplification of individual loci, which are especially prominent for low-efficiency origins.

In cis effects. Next, we investigated origins whose amplification levels could not be explained merely by their intrinsic efficiency. A relevant example is given in Figure 5E; from the distribution it is clear that, although Ori I-287 has a very low efficiency, its amplification levels are much higher than expected. A closer examination of the neighboring origins reveals that its left-flanking origin (Ori I-288, at a genomic distance of 17424 bp) is one of the most efficient in the genome, with a firing efficiency of 73%; their copy number levels are strongly correlated ($\rho = 0.98$). To further investigate this, we computed correlation coefficients across copy number levels of all origins in the genome (whole genome in Supplementary Figure S8 A and B, zoom in Chromosome III—end in Figure 5F). This analysis indicated strong correlations between adjacent origins, pointing to *in cis* effects. To better understand the extent of this effect, we computed correlations between copy number levels of the 10 peaks with the highest amplification and their neighborhood (Figure 5G). Since our previous analysis showed that fork speed affects the extent of passive re-replication, we also computed correlations using the simulations with a fork speed of 1 kb/min and 3 kb/min. The results reveal that copy numbers of each central amplification origin are significantly positively correlated with the ones of its right and left flanking up to a distance of 0.1 megabases; it is also clear that as speed increases, the extent of positive correlation increases

as well and for fork speed equal to 3 kb/min it reaches a distance of 0.5 megabases.

We then asked how the firing activity of individual origins is affected by the activity of its neighbors. We focused our analysis on Ori I-287 and Ori I-40, two origins that share the same low efficiency (9%), but Ori I-287 has an immediate neighbor with high efficiency (73%) whereas Ori I-40 does not. We observed that Ori I-40 does not fire the majority of times, whereas Ori I-287 appears more active and occasionally fires even more than five times (Figure 5H). We then followed the same methodology as above and computed the correlation coefficient between the number of fires of different origins across the genome (Figure 5I). This analysis indicated that indeed local effects exist, suggesting that the more times an origin fires, the more will its neighbors fire as well. We notice that the firing events of each central amplification origin are significantly positively correlated with the ones of its immediate right and left flanking neighbors, however, this time the correlation spans a smaller region, drops sharply with distance from the central origin and does not appear affected by fork speed (Figure 5J). These findings indicate that, in addition to passive re-replication, *in cis* effects between adjacent origin locations are also implicitly attributed to increased total firing activity of weak origins located close to strong origins. Early firing of a strong origin will increase the newly born copies of a nearby weak origin, facilitating its re-firing.

In trans effects. To explore the variability of the re-replication process genome-wide, we performed a principal component analysis of the genome-wide amplification profiles of 100 simulations at 16*C* and visualized the results as a biplot of the first two principal components (Figure 6A), where dots correspond to simulations and vectors indicate the PCA loadings, i.e. the correlation of each origin to the unit-scaled first two principal components. From this it becomes clear that a large amount of the variability in the simulations is dominated by two different origins of Chromosome III (Ori III-11 and Ori III-118). Specifically, the first and second principal component correlate strongly with Ori III-118 and Ori III-11, respectively, while Ori III-11 additionally appears to correlate negatively with principal component 1.

To further explore how specific origins may affect genome-wide amplification profiles, we clustered profiles using *k*-means clustering, estimated the optimal *k* using the Gap statistic (52) and the stability of the clustering using the Adjusted Rand Index (ARI) (53) (details in 'Materials and Methods' section). For 100 simulations at 16*C* an optimal number of $k = 3$ clusters was identified, and the cluster assignments were very consistent across 100 random initializations, with a mean ARI of 0.95 (standard deviation = 0.05). These clusters correspond to 3 groups of simulations characterized by different patterns of re-replication at a genome level (Figure 6B). The clusters appear to be dominated by the amplification of origins Ori III-11 and Ori III-118 in a mutually exclusive manner: either one of the two origins is amplified (clusters 1 and 3) or they are both relatively low (cluster 2). Indeed, as shown in Figure 6C, levels of amplification of Ori III-11 and Ori III-118 are negatively correlated in individual simulations ($\rho = -0.3$, *P*-value = 0.0025) and characterize the three clusters. Though re-replication at OriIII-118 may be overestimated in simulations in comparison to experimental re-replication data (Figure 2), our analysis indicates that highly efficient origins can interfere with each other during re-replication even when far apart. Taken together, these findings indicate *in trans* effects within the genome.

We next examined the same simulations at a DNA content of 2*C* (Figure 6D, same ordering as in Figure 6B). We observe that while specific amplification regions are starting to emerge, the re-replication levels for the majority of the genome are around the genome mean, amplification occurs in random regions along the genome and the process is governed by a high degree of variability. At the same time the difference between single-cell profiles of the previously identified clusters is not noticeable. To validate this, we went on to independently cluster the data and estimated an optimal *k* of only 1 cluster. Forcing *k* equal to 3 and estimating the stability across 100 random initializations indicated close to random cluster assignment between different runs (ARI = 0.36 ± 0.15). Comparing the clusters found for 2*C* and 16*C* also indicated very low agreement (ARI = 0.34 ± 0.10). Last, the correlation coefficient between Ori III-11 and Ori III-118 for 2C is now non-significant ($\rho = -0.09$, *P*-value = 0.36).

We conclude that re-replication is initially characterized by a high degree of randomness, while *in trans* effects become evident as the re-replication process progresses, leading to preferred genome-wide patterns of re-replication at high DNA content. These are dominated by a small number of high activity origins, whose amplification to high levels is mutually exclusive.

DISCUSSION

A stochastic hybrid model of DNA re-replication

In this work a stochastic hybrid model of DNA re-replication was presented, developed by refining existing work of normal DNA replication so that it allows for origin re-firing. The model accurately portrays the interplay between discrete dynamics, associated with different origin states, continuous dynamics, associated with the movement of the replication forks, and stochasticity, associated with random firing and re-firing events. Transitions between discrete states depend on both the continuous and stochastic dynamics of the system, such as firing events or merging of neighboring forks. In addition, two automatic transitions, specific to the re-replication case, are incorporated in the model; when an origin fires or is passively replicated its descendants automatically fall into the pre-replicative state and can potentially fire or be passively replicated again.

Using input data from experimentally determined origin locations and intrinsic firing efficiencies from fission yeast, the model allows the simulation of re-replication along the complete fission yeast genome and thus the exploration of re-replication kinetics genome-wide. Two alternative variations of the model have been implemented, depending on how the firing propensities of the newly born origins are assigned. In the base-case model variation (LF model), the total system propensity is kept constant and continuously redistributed to all existing and newly born origins. The

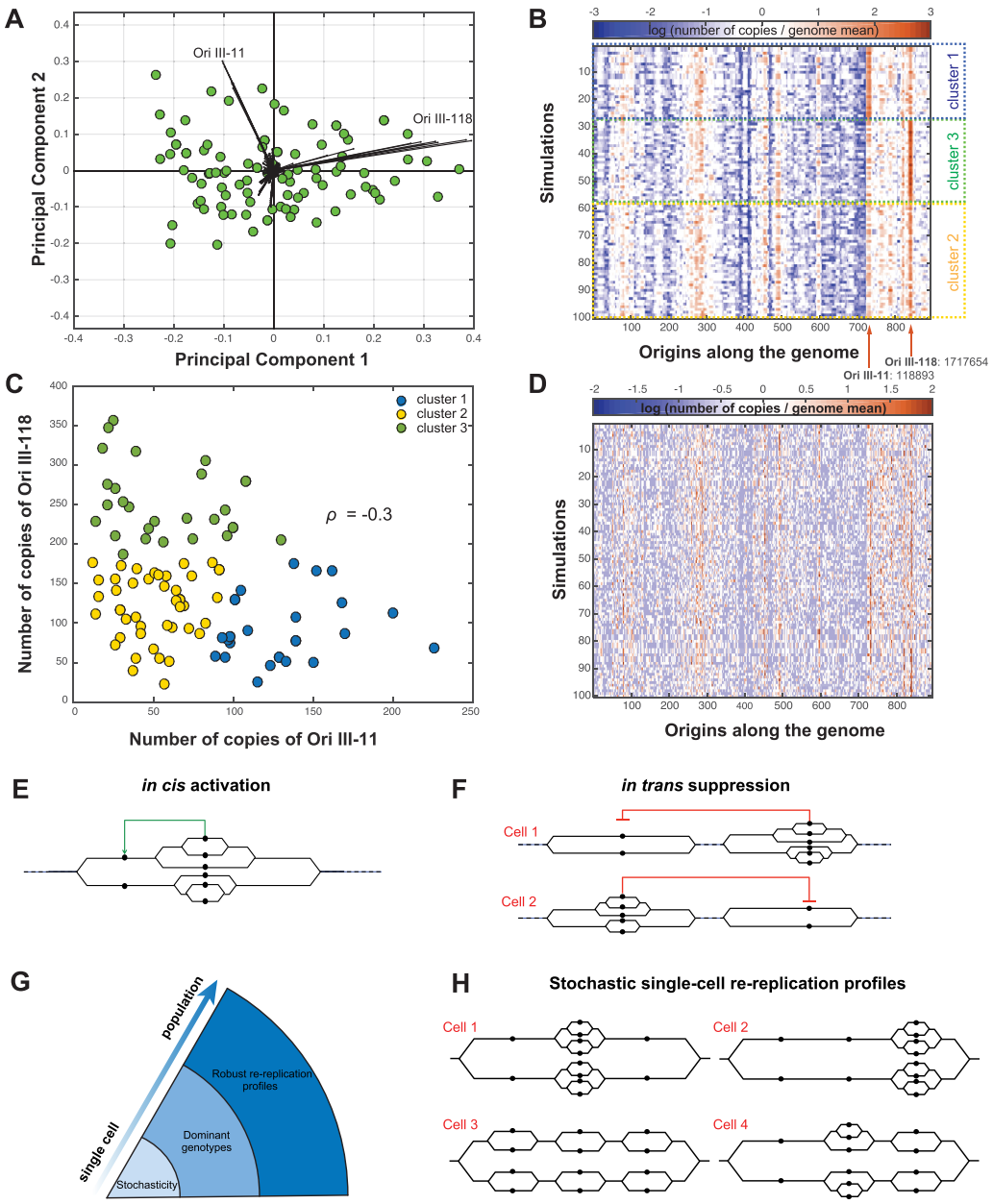

Figure 6. Analysis of *in silico* data at a whole genome level points to *in trans* effects within the genome. (**A**) Variability of copy number levels genome-wide is governed by prominent origins. Results of a PCA analysis of the *in silico* copy number data, shown as a biplot of the first two principal components. Dots correspond to simulations and black vectors expose each origin's contribution to the first two components, both in terms of magnitude and direction (marked here for the two most prominent ones). (**B**) Heatmap of DNA content (rows: simulations, columns: origins) for 100 simulations at 16***C*** after clustering with a k-means algorithm and $k = 3$. Color indicates DNA amplification levels, expressed as the log ratio of individual versus genome mean number of copies. Identified clusters are marked with different colors. (**C**) Scatterplot of number of copies for origins Ori III-11 and Ori III-118 shows a negative correlation ($\rho = -0.4$). Colors correspond to simulations belonging to each of the three clusters identified in B. (**D**) Evolution of re-replication over time. Heatmap of DNA content for simulations of (B) at an earlier DNA content of 2***C*** shows no cluster-specific patterns at a low-re-replication context. (**E**) Underlying characteristics of DNA re-replication. *In cis* effects between adjacent loci. Passive re-replication of inactive origins from their efficient neighbors leads to increased copy numbers and implicitly increases their firing activity. (**F**) *In trans* effects between distant loci. Increased amplification of one locus leads to in trans suppression of a distant locus. (**G**) Emerging properties of DNA re-replication, depending on the level of analysis. (**H**) *In silico* re-replication profiles. Simulation results reveal many possible genotypes within a population, shown here in a schematic view for three hypothetical origins. Although the total DNA content is the same in all four single cells, individual copy number levels vary greatly.

base-case model therefore assumes the presence of a limiting factor, which restrains origins firing system-wide. In (39) it was shown that, for normal replication, redistribution of a limiting factor increases the efficiency of remaining origins and may help explain away the so-called random gap problem (64). During re-replication, such a limiting factor could act at the licensing step (for example limited ORC, Cdt1, Cdc6/18 or MCM levels), at the firing level (for example limiting Cdc45 or DDK kinase levels) (44,65–67), or both. The model in its current instantiation does not discriminate between these two possibilities, it could however be easily modified to explicitly model the licensing and firing events separately. In the alternative variation (UF model), offspring origins inherit the same firing propensity as the parent, and, as the total number of origins increases exponentially, the total system firing propensity will also increase. This variation simulates unlimited re-replication.

Parameters affecting re-replication dynamics

Comparison of *in silico* data for both model variations and experimentation with different values of the model inputs has permitted insight into the model parameters affecting re-replication dynamics. Our analysis indicated that the simulated re-replication completion times were consistent with experimental observations when using the model variation with limiting factor and a fork speed of 0.5 kb/min. This indicates that, as expected, fork speed is slower in re-replication that in normal S-phase, where experimental estimates in yeast vary between 1.6 and 3 kb/min. Since re-replication is a non-physiological process, differences in fork speed could be attributed to various mechanisms, such as activation of checkpoint proteins that stall the forks, impediments caused by fork collisions (68) and limitations in the amounts of various necessary substrates like dNTPs (DNA building blocks), activation factors etc. In the current instantiation, we have modeled fork progression as a deterministic event, and we have assumed a constant fork speed across the genome and through time. All these assumptions can be relaxed. Our model can accommodate different fork speeds at different locations or different points in time as re-replication progresses, to simulate for example difficult to replicate regions or scarcity of dNTPs. In addition, fork movement can be modeled as a stochastic process (69), permitting stochastic fork slowing or arrest. Such an instantiation would be highly relevant for re-replication, where fork slowing or arrest would lead to head to tail fork collisions resulting in double strand breaks—an event observed experimentally (70).

Further exploration using different model variants indicated that when no limiting factor is assumed, the rate of increase in DNA content far exceeds experimental observations. At the same time, in the model variation with limiting factor, re-replication dynamics are dominated by passive replication instead of firing events, whereas in the variation without limiting factor, firing and passive replication contribute equally to the increase in DNA content. Sensitivity analysis using different values of fork speed showed that, when fork speed is decreased, more time is needed to reach the desired DNA content. At the same time an apparent trade-off between fork speed and firing events was noticed, since faster forks resulted in less firing and allowed passive replication to dominate the increase in DNA content.

Genome-wide profile of re-replication

Analyzing the simulated data at a population level, it is clear that re-replication is non-homogeneous along the genome, as specific regions are preferably amplified and appear as emerging peaks above the genome mean, whereas others appear dormant and under-represented, an observation that is in accordance with existing experimental findings (12,13). Although the model is stochastic, the most highly amplified regions appear to be highly robust with respect to different model variations or different values of the fork speed. By comparing the simulated versus the experimental amplification profiles, we observe that overall the simulated data reproduce the experimental re-replication pattern on a whole-genome scale, validating our approach. The best-fitting parameter set proved to be when using the model variation that assumes the existence of limiting factor and a fork speed of 0.5 kb/min, reconfirming the previous findings.

Our analysis showed that most highly amplified regions in experimental data are predicted when using the simulations, with some inconsistencies attributed to minor linear shifts or differences in intensity. Striking differences regard specific regions, with the most prominent ones being the subtelomeres, regions highly amplified in experimental data. This difference could be attributed to location-specific mechanisms, such as the suppression of the telomeric origins in normal DNA replication by telomere-associated proteins Rif1 and Taz1. Distorted nuclear architecture during re-replication or limiting abundance of Rif1/Taz1 could lead to the subtelomeric origins escaping their normal control mechanism and getting amplified above the levels that are expected from their experimentally determined mitotic efficiencies (31).

Factors that affect amplification levels of individual loci

Analysis of simulated data at a single-cell level permits a more detailed insight into DNA re-replication. Amplification levels of individual loci are primarily affected by intrinsic properties, since copy numbers were found to be highly correlated with firing efficiencies. At the same time, we found that individual origins are able to act *in cis* and amplify the copy numbers of their neighbors. Positive correlation between amplification levels of adjacent loci is primarily attributed to passive re-replication, as forks emanating from the firing origins to both directions passively replicate the left and right flanking origins. At the same time, as the forks create new copies of the passively replicated origins, these newly born copies can potentially fire again, thus increasing the overall probability of firing events from the neighboring origins. This means that *in cis* elements contribute to amplified copy numbers not only directly by passive re-replication, but also implicitly through increasing the probability that their neighbors will fire, due to their increased copy number (Figure 6E). This type of positive correlation between adjoining regions is a key characteristic of re-replication and serves as a mechanism for indirect amplification of individual loci.

At the same time genome-wide analysis of the single-cell re-replication profiles revealed groups of similarly amplified simulations, characterized by different patterns of re-replication. These patterns were characterized by amplification of specific regions, residing in non-proximal locations across the genome that appeared to exist in opposition. The amplification levels of these loci were found to be negatively correlated and allowed for a clear separation of the clusters. These findings point to *in trans* interactions between distant regions within the genome, suggesting a mechanism for the suppression of the amplification levels of individual loci (Figure 6F). Such *in trans* negative regulation of distant origins could be explained by competition for the same limiting factor; high-level amplification of a given locus recruits high levels of the limiting factor, indirectly inhibiting firing of other genomic regions.

Emerging properties of re-replication, revealed by different levels of analysis

Depending on the level of analysis (from the single-cell toward the population level), different properties of re-replication are revealed. At the single-cell level a large degree of heterogeneity is observed, not only in the variations of individual loci among the population, but also in the variability of single-cell profiles. When clustered, different amplification patterns are recognized in these profiles, corresponding to dominant genotypes within the population. Last, at a population level, re-replication appears highly robust and amplification hotspots appear independent of changes in parameters (Figure 6G). In conclusion, heterogeneity and robustness appear as key players in the re-replicating process that co-exist and act in parallel. This implies that experimental observations of re-replication, based on population-level data, possibly mask the underlying variability in the behavior of single-cells in a population.

Cell-to-cell heterogeneity leads to genome plasticity

Stochasticity lies at the heart of re-replication: it gives rise to heterogeneous single-cell profiles that correspond to diverse genotypes within the population. Although amplification levels of individual loci within the population are affected by intrinsic properties, *cis*- and *trans*-acting elements as mentioned above, great variations in copy numbers of individual loci within a population are revealed, evident by skewed distributions with long tails. As each simulation of the stochastic model corresponds to a single cell in a population, each simulated re-replication profile portrays a unique sequence of firing and re-firing events and corresponds to different genotypes within the population (Figure 6H). By exploiting this property of the model, our analysis showed that, although re-replication profiles at the population level are robust, at the single-cell level they are heterogeneous and can deviate significantly from the mean. Early firing of an origin in a given cell will increase the probability of a second firing event in the same locus (as more strands are born), leading to a positive feedback loop that will amplify different loci in different cells.

At the same time, using a population size of 100 simulations, the majority of the genome was found to be amplified above the genome mean at least once. This suggests that even regions of low efficiency can potentially be amplified, and that re-replication can, with varying probability, occur anywhere in the genome and generate many diverse genotypes within a population. If we consider that a small colony of yeast cells contains millions of individual cells, it becomes apparent that re-replication can lead to the appearance of amplification events in a variety of chromosomal regions or combinations of regions. These observations indicate that cell-to-cell variability is inherent in re-replication and can lead to a high degree of genome plasticity. By tracking the evolution of single-cell profiles from a low to a high re-replication context, we found that cell-to-cell variability is more prominent at the onset of re-replication, when single-cell profiles appear highly stochastic in nature. As DNA content increases, cell-to-cell variability is less apparent and specific amplification regions dominate the process. At the same time, distinct patterns gradually emerge, representing dominant genotypes within the population that appear to act antagonistically.

Genome plasticity and possible implications for oncogenesis

Variations in the number of copies of specific genomic loci have long been implicated in the initiation and progression of cancer. For example, oncogenes and genes conferring resistance to drugs have been shown to be frequently amplified in various cancers (26,71–74). Re-replication, and the resulting increase in the copies of specific genomic loci, could be a mechanism leading to gene amplification (18,75). Studies using cancer genome data correlate replication timing with mutation rates during cancer and suggest that early replication is correlated with gene amplifications whereas late replication with copy number losses (76–78). An overwhelming amount of experimental evidence supports a high level of heterogeneity in cancer cell populations (79–82). Recent studies using next-generation sequencing have revealed that cancer genomes evolve dynamically through different trajectories even within the same tumor (83–86). In our work, we have demonstrated *in silico* that re-replication can promote genome plasticity, by generating many diverse genotypes within a population. In cells, incorporation of re-replicating regions into the genome could create site-specific copy gains leading to heterogeneous phenotypes, with potentially desired properties. In a context of natural selection, re-replication may offer a great evolutionary advantage in cells that have lost their normal replication controls, by enabling them to dynamically obtain desired phenotypes and adapt to their environment. Future work will allow such mechanisms to be investigated *in vivo*.

In summary, we have developed the first mathematical model of DNA re-replication, and extensively simulated it for different hypotheses and model parameters across the complete fission yeast genome. Our *in silico* analysis has elucidated the basic principles that govern DNA re-replication, and indicated how these are manifested depending on the level of analysis: although at the single-cell level re-replication is stochastic and any genomic region is susceptible to amplification, genome-wide patterns of amplification at the population level are robust to different hypotheses and model parameters. These observations highlight that heterogeneity and robustness are emerging and

non-contradictory characteristics of DNA re-replication. Importantly, by demonstrating the link between DNA re-replication and genome plasticity, our work may have broad implications for better understanding the onset of genomic instability and cancer evolution.

DATA AVAILABILITY

The model's source code, all generated data and extensively documented figure-generating scripts are available under an open-source license on GitHub: https://github.com/rapsoman/DNA_Rereplication.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We acknowledge the contribution of Dr Konstantinos Koutroumpas in the design and implementation of the mathematical model. We thank the Advanced Light Microscopy Facility of the University of Patras for support with imaging.
Author contributions: J.L. and Z.L. conceived and supervised the study. J.L. and Z.L. designed the mathematical model. M.A.R. implemented the model, ran all simulations, collected and analyzed *in silico* data. M.A.R., J.L. and Z.L. interpreted *in silico* data. S.M., M.R.G., P.N., N.N.G., S.T. and Z.L. designed, performed and/or analyzed biological experiments. M.A.R., J.L. and Z.L. wrote the manuscript, with input from all authors. All authors read and approved the final manuscript.

FUNDING

European Research Council [ERC-StG 281851, ERC-PoC 755284]; State Scholarships Foundation of Greece, Short-term Fellowship (to M.A.R.); State Scholarships Foundation of Greece, Ph.D. Fellowship (to S.M.). This work was supported by the project "Bioimaging-GR" (MIS 5002755), implemented under the Action "Reinforcement of the Research and Innovation Infrastructure", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the EU. Funding for open access charge: Bioimaging-GR.
Conflict of interest statement. None declared.

REFERENCES

1. Symeonidou,I.E., Taraviras,S. and Lygerou,Z. (2012) Control over DNA replication in time and space. *FEBS Lett.* **586**, 2803–2812.
2. Siddiqui,K., On,K.F. and Diffley,J.F.X. (2013) Regulating DNA replication in Eukarya. *Cold Spring Harb. Perspect. Biol.*, **5**, a012930.
3. Rhind,N. and Gilbert,D.M. (2013) DNA replication timing. *Cold Spring Harb. Perspect. Biol.*, **5**, a010132.
4. Rhind,N., Yang,S.C.-H. and Bechhoefer,J. (2010) Reconciling stochastic origin firing with defined replication timing. *Chromosome Res.*, **18**, 35–43.
5. Fragkos,M., Ganier,O., Coulombe,P. and Méchali,M. (2015) DNA replication origin activation in space and time. *Nat. Rev. Mol. Cell Biol.*, **16**, 360–374.
6. Legouras,I., Xouri,G., Dimopoulos,S., Lygeros,J. and Lygerou,Z. (2006) DNA replication in the fission yeast: robustness in the face of uncertainty. *Yeast*, **23**, 951–962.
7. Nathanailidou,P., Taraviras,S. and Lygerou,Z. (2020) Chromatin and nuclear architecture: shaping DNA replication in 3D. *Trends Genet.*, **36**, 967–980.
8. Nishitani,N. and Lygerou,Z. (2004) DNA replication licensing. *Front. Biosci.*, **9**, 2115–2132.
9. Parker,M.W., Botchan,M.R. and Berger,J.M. (2017) Mechanisms and regulation of DNA replication initiation in eukaryotes. *Crit. Rev. Biochem. Mol. Biol.*, **52**, 107–144.
10. Blow,J.J. and Dutta,A. (2005) Preventing re-replication of chromosomal DNA. *Nat. Rev. Mol. Cell Biol.*, **6**, 476–486.
11. Nishitani,H. and Nurse,P. (1995) p65cdc18 plays a major role controlling the initiation of DNA replication in fission yeast. *Cell*, **83**, 397–405.
12. Mickle,K.L., Oliva,A., Huberman,J.A. and Leatherwood,J. (2007) Checkpoint effects and telomere amplification during DNA re-replication in fission yeast. *BMC Mol. Biol.*, **8**, 119.
13. Kiang,L., Heichinger,C., Watt,S., Bähler,J. and Nurse,P. (2010) Specific replication origins promote DNA amplification in fission yeast. *J. Cell Sci.*, **123**, 3047–3051.
14. Nishitani,H., Lygerou,Z., Nishimoto,T. and Nurse,P. (2000) The Cdt1 protein is required to license DNA for replication in fission yeast. *Nature*, **404**, 625–628.
15. Yanow,S.K., Lygerou,Z. and Nurse,P. (2001) Expression of Cdc18/Cdc6 and Cdt1 during G2 phase induces initiation of DNA replication. *EMBO J.*, **20**, 4648–4656.
16. Green,B.M., Morreale,R.J., Özaydin,B., DeRisi,J.L. and Li,J.J. (2006) Genome-wide mapping of DNA synthesis in *Saccharomyces cerevisiae* reveals that mechanisms preventing reinitiation of DNA replication are not redundant. *Mol. Biol. Cell*, **17**, 2401–2414.
17. Tanny,R.E., MacAlpine,D.M., Blitzblau,H.G. and Bell,S.P. (2006) Genome-wide analysis of re-replication reveals inhibitory controls that target multiple stages of replication initiation. *Mol. Biol. Cell*, **17**, 2415–2423.
18. Green,B.M., Finn,K.J. and Li,J.J. (2010) Loss of DNA replication control is a potent inducer of gene amplification. *Science*, **329**, 943–946.
19. Finn,K.J. and Li,J.J. (2013) Single-stranded annealing induced by re-initiation of replication origins provides a novel and efficient mechanism for generating copy number expansion via non-allelic homologous recombination. *PLos Genet.*, **9**, e1003192.
20. Bui,D.T. and Li,J.J. (2019) DNA rereplication is susceptible to nucleotide-level mutagenesis. *Genetics*, **212**, 445–460.
21. Vaziri,C., Saxena,S., Jeon,Y., Lee,C., Murata,K., Machida,Y., Wagle,N., Hwang,D.S. and Dutta,A. (2003) A p53-dependent checkpoint pathway prevents rereplication. *Mol. Cell*, **11**, 997–1008.
22. Karakaidos,P., Taraviras,S., Vassiliou,L.V., Zacharatos,P., Kastrinakis,N.G., Kougiou,D., Kouloukoussa,M., Nishitani,H., Papavassiliou,A.G., Lygerou,Z. *et al.* (2004) Overexpression of the replication licensing regulators hCdt1 and hCdc6 characterizes a subset of non-small-cell lung carcinomas: synergistic effect with mutant p53 on tumor growth and chromosomal instability—evidence of E2F-1 transcriptional control over hCdt1. *Am. J. Pathol.*, **165**, 1351–1365.
23. Liontos,M., Koutsami,M., Sideridou,M., Evangelou,K., Kletsas,D., Levy,B., Kotsinas,A., Nahum,O., Zoumpourlis,V., Kouloukoussa,M. *et al.* (2007) Deregulated overexpression of hCdt1 and hCdc6 promotes malignant behavior. *Cancer Res.*, **67**, 10899–10909.
24. Champeris-Tsaniras,S., Kanellakis,N., Symeonidou,I.E., Nikolopoulou,P., Lygerou,Z. and Taraviras,S. (2014) Licensing of DNA replication, cancer, pluripotency and differentiation: an interlinked world? *Semin. Cell Dev. Biol.*, **30**, 174–180.
25. Gaillard,H., García-Muse,T. and Aguilera,A. (2015) Replication stress and cancer. *Nat. Rev. Cancer*, **15**, 276–289.
26. Hills,S.A. and Diffley,J.F.X. (2014) DNA replication and oncogene-induced replicative stress. *Curr. Biol.*, **24**, R435–R444.
27. Hook,S.S., Lin,J.J. and Dutta,A. (2007) Mechanisms to control rereplication and implications for cancer. *Curr. Opin. Cell Biol.*, **19**, 663–671.
28. Nowak,M.A., Komarova,N.L., Sengupta,A., Jallepalli,P.V., Shih,I.M., Vogelstein,B. and Lengauer,C. (2002) The role of

chromosomal instability in tumor initiation. *Proc. Natl Acad. Sci. U.S.A.*, **99**, 16226–16231.
29. Petropoulos,M., Tsaniras,S.C., Taraviras,S. and Lygerou,Z. (2019) Replication licensing aberrations, replication stress, and genomic instability. *Trends Biochem. Sci.*, **44**, 752–764.
30. Petropoulou,C., Kotantaki,P., Karamitros,D. and Taraviras,S. (2008) Cdt1 and Geminin in cancer: markers or triggers of malignant transformation? *Front. Biosci.*, **13**, 4485–4494.
31. Heichinger,C., Penkett,C.J., Bähler,J. and Nurse,P. (2006) Genome-wide characterization of fission yeast DNA replication origins. *EMBO J.*, **25**, 5171–5179.
32. Blow,J.J. and Ge,X.Q. (2009) A model for DNA replication showing how dormant origins safeguard against replication fork failure. *EMBO Rep.*, **10**, 406–412.
33. Brümmer,A., Salazar,C., Zinzalla,V., Alberghina,L. and Höfer,T. (2010) Mathematical modelling of DNA replication reveals a trade-off between coherence of origin activation and robustness against rereplication. *PLoS Comput. Biol.*, **6**, e1000783.
34. Gauthier,M.G. and Bechhoefer,J. (2009) Control of DNA replication by anomalous reaction-diffusion kinetics. *Phys. Rev. Lett.*, **102**, 158104.
35. Gauthier,M.G., Norio,P. and Bechhoefer,J. (2012) Modeling inhomogeneous DNA replication kinetics. *PLoS One*, **7**, e32053.
36. Gispan,A., Carmi,M. and Barkai,N. (2017) Model-based analysis of DNA replication profiles: predicting replication fork velocity and initiation rate by profiling free-cycling cells. *Genome Res.*, **27**, 310–319.
37. Goldar,A., Labit,H., Marheineke,K. and Hyrien,O. (2008) A dynamic stochastic model for DNA replication initiation in early embryos. *PLoS One*, **3**, e2919.
38. Kelly,T. and Callegari,A.J. (2019) Dynamics of DNA replication in a eukaryotic cell. *Proc. Natl Acad. Sci. U.S.A.*, **116**, 4973–4982
39. Lygeros,J., Koutroumpas,K., Dimopoulos,S., Legouras,I., Kouretas,P., Heichinger,C., Nurse,P. and Lygerou,Z. (2008) Stochastic hybrid modeling of DNA replication across a complete genome. *Proc. Natl Acad. Sci. U.S.A.*, **105**, 12295–12300.
40. Al Mamun,M., Albergante,L., Moreno,A., Carrington,J.T., Blow,J.J. and Newman,T.J. (2016) Inevitability and containment of replication errors for eukaryotic genome lengths spanning megabase to gigabase. *Proc. Natl Acad. Sci. U.S.A.*, **113**, E5765–E5774.
41. de Moura,A.P.S., Retkute,R., Hawkins,M. and Nieduszynski,C.A. (2010) Mathematical modelling of whole chromosome replication. *Nucleic Acids Res.*, **38**, 5623–5633.
42. Demczuk,A., Gauthier,M.G., Veras,I., Kosiyatrakul,S., Schildkraut,C.L., Busslinger,M., Bechhoefer,J. and Norio,P. (2012) Regulation of DNA replication within the immunoglobulin heavy-chain locus during B cell commitment. *PLoS Biol.*, **10**, e1001360.
43. Kaykov,A. and Nurse,P. (2015) The spatial and temporal organization of origin firing during the S-phase of fission yeast. *Genome Res.*, **25**, 391–401.
44. Patel,P.K., Kommajosyula,N., Rosebrock,A., Bensimon,A., Leatherwood,J., Bechhoefer,J. and Rhind,N. (2008) The Hsk1(Cdc7) replication kinase regulates origin efficiency. *Mol. Biol. Cell*, **19**, 5550–5558.
45. Richardson,C.D. and Li,J.J. (2014) Regulatory mechanisms that prevent re-initiation of DNA replication can be locally modulated at origins by nearby sequence elements. *PLos Genet.*, **10**, e1004358.
46. Menzel,J., Tatman,P. and Black,J.C. (2020) Isolation and analysis of rereplicated DNA by Rerep-Seq. *Nucleic Acids Res.*, **48**, e58.
47. Cassandras,C.G. and Lygeros,J. (2007) In: *Stochastic Hybrid Systems*. CRC Press, Boca Raton.
48. Kouretas,P., Koutroumpas,K., Lygeros,J. and Lygerou,Z. (2006) Stochastic hybrid modeling of biochemical processes. In: Cassandras,C.G. and Lygeros,J. (eds). *Stochastic Hybrid Systems*. CRC Press, Boca Raton, pp. 221–248.
49. Cinquemani,E., Milias-Argeitis,A., Summers,S. and Lygeros,J. (2008) Stochastic dynamics of genetic networks: modelling and parameter identification. *Bioinformatics*, **24**, 2748–2754.
50. Rapsomaniki,M.A., Cinquemani,E., Giakoumakis,N.N., Kotsantis,P., Lygeros,J. and Lygerou,Z. (2015) Inference of protein kinetics by stochastic modeling and simulation of fluorescence recovery after photobleaching experiments. *Bioinformatics*, **31**, 355–362.
51. Daigaku,Y., Keszthelyi,A., Müller,C.A., Miyabe,I., Brooks,T., Retkute,R., Hubank,M., Nieduszynski,C.A. and Carr,A.M. (2015) A global profile of replicative polymerase usage. *Nat. Struct. Mol. Biol.*, **22**, 192–198.
52. Tibshirani,R., Walther,G. and Hastie,T. (2001) Estimating the number of clusters in a data set via the gap statistic. *J. Royal Stat. Soc.*, **63**, 411–423.
53. Hubert,L. and Arabie,P. (1985) Comparing partitions. *J. Classification*, **2**, 193–218.
54. Petrova,B., Dehler,S., Kruitwagen,T., Hériché,J.K., Miura,K. and Haering,C.H. (2013) Quantitative analysis of chromosome condensation in fission yeast. *Mol. Cell. Biol.*, **33**, 984–998.
55. Baum,B. (1998) Cdc18 transcription and proteolysis couple S phase to passage through mitosis. *EMBO J.*, **17**, 5689–5698.
56. Koutroumpas,K. and Lygeros,J. (2011) Modeling and analysis of DNA replication. *Automatica*, **47**, 1156–1164.
57. Raghuraman,M.K., Winzeler,E.A., Collingwood,D., Hunt,S., Wodicka,L., Conway,A., Lockhart,D.J., Davis,R.W., Brewer,B.J. and Fangman,W.L. (2001) Replication dynamics of the yeast genome. *Science*, **294**, 115–121.
58. Yabuki,N., Terashima,H. and Kitada,K. (2002) Mapping of early firing origins on a replication profile of budding yeast. *Genes Cells*, **7**, 781–789.
59. Sekedat,M.D., Fenyö,D., Rogers,R.S., Tackett,A.J., Aitchison,J.D. and Chait,B.T. (2010) GINS motion reveals replication fork progression is remarkably uniform throughout the yeast genome. *Mol. Syst. Biol.*, **6**, 353.
60. Duzdevich,D., Warner,M.D., Ticau,S., Ivica,N.A., Bell,S.P. and Greene,E.C. (2015) The dynamics of eukaryotic replication initiation: origin specificity, licensing, and firing at the single-molecule level. *Mol. Cell*, **58**, 483–494.
61. Rhind,N. (2006) DNA replication timing: random thoughts about origin firing. *Nat. Cell Biol.*, **8**, 1313–1316.
62. Kitamura,E., Blow,J.J. and Tanaka,T.U. (2006) Live-cell imaging reveals replication of individual replicons in eukaryotic replication factories. *Cell*, **125**, 1297–1308.
63. Greenwood,E., Nishitani,H. and Nurse,P. (1998) Cdc18p can block mitosis by two independent mechanisms. *J. Cell Sci.*, **20**, 3101–3108.
64. Mantiero,D., Mackenzie,A., Donaldson,A. and Zegerman,P. (2011) Limiting replication initiation factors execute the temporal programme of origin firing in budding yeast. *EMBO J.*, **30**, 4805–4814.
65. Aparicio,O.M. (2013) Location, location, location: it's all in the timing for replication origins. *Genes Dev.*, **27**, 117–128.
66. Wu,P.Y.J. and Nurse,P. (2009) Establishing the program of origin firing during S phase in fission yeast. *Cell*, **136**, 852–864.
67. Hyrien,O., Marheineke,K. and Goldar,A. (2003) Paradoxes of eukaryotic DNA replication: MCM proteins and the random completion problem. *Bioessays*, **25**, 116–125.
68. Alexander,J.L. and Orr-Weaver,T.L. (2016) Replication fork instability and the consequences of fork collisions from rereplication. *Genes Dev.*, **30**, 2241–2252.
69. Boemo,M.A., Cardelli,L. and Nieduszynski,C.A. (2020) The Beacon Calculus: a formal method for the flexible and concise modelling of biological systems. *PLoS Comput. Biol.*, **16**, e1007651.
70. Davidson,I.F., Li,A. and Blow,J.J. (2006) Deregulated replication licensing causes DNA fragmentation consistent with Head-to-Tail fork collision. *Mol. Cell*, **24**, 433–443.
71. Masood,S. and Bui,M.M. (2002) Prognostic and predictive value of HER2/neu oncogene in breast cancer. *Microsc. Res. Tech.*, **59**, 102–108.
72. Ross,J.S., Fletcher,J.A., Bloom,K.J., Linette,G.P., Stec,J., Symmans,W.F., Pusztai,L. and Hortobagyi,G.N. (2004) Targeted therapy in breast cancer: the HER-2/neu gene and protein. *Mol. Cell Proteomics*, **3**, 379–398.
73. Santarius,T., Shipley,J., Brewer,D., Stratton,M.R. and Cooper,C.S. (2010) A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer*, **10**, 59–64.
74. Schmitt,M.W., Loeb,L.A. and Salk,J.J. (2016) The influence of subclonal resistance mutations on targeted cancer therapy. *Nat. Rev. Clin. Oncol.*, **13**, 335–347.
75. Black,J.C., Manning,A.L., Van Rechem,C., Kim,J., Ladd,B., Cho,J., Pineda,C.M., Murphy,N., Daniels,D.L., Montagna,C. *et al.* (2013)

KDM4A lysine demethylase induces site-specific copy gain and rereplication of regions amplified in tumors. *Cell*, **154**, 541–555.
76. De,S. and Michor,F. (2011) DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat. Biotech.*, **29**, 1103–1108.
77. Sima,J. and Gilbert,D.M. (2014) Complex correlations: replication timing and mutational landscapes during cancer and genome evolution. *Curr. Opin. Genet. Dev.*, **25**, 93–100.
78. Miotto,B., Ji,Z. and Struhl,K. (2016) Selectivity of ORC binding sites and the relation to replication timing, fragile sites, and deletions in cancers. *Proc. Natl Acad. Sci. U.S.A.*, **113**, E4810–E4819.
79. Greaves,M. and Maley,C.C. (2012) Clonal evolution in cancer. *Nature*, **481**, 306–313.
80. Marusyk,A., Almendro,V. and Polyak,K. (2012) Intra-tumour heterogeneity: a looking glass for cancer? *Nat. Rev. Cancer*, **12**, 323–334.
81. Burrell,R.A. and Swanton,C. (2014) The evolution of the unstable cancer genome. *Curr. Opin. Gen. Dev.*, **24**, 61–67.
82. Zhang,C.Z. and Pellman,D. (2016) From mutational mechanisms in single cells to mutational patterns in cancer genomes. *Cold Spring Harb. Symp. Quant. Biol.*, **11**, 027623.
83. Almendro,V., Cheng,Y.K., Randles,A., Itzkovitz,S., Marusyk,A., Ametller,E., Gonzalez-Farre,X., Muñoz,M., Russnes,H.G., Helland,A. *et al.* (2014) Inference of tumor evolution during chemotherapy by computational modeling and in situ analysis of cellular diversity for genetic and phenotypic features. *Cell Rep.*, **6**, 514–527.
84. Marusyk,A., Tabassum,D.P., Altrock,P.M., Almendro,V., Michor,F. and Polyak,K. (2014) Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature*, **514**, 54–58.
85. Gao,R., Davis,A., McDonald,T.O., Sei,E., Shi,X., Wang,Y., Tsai,P.C., Casasent,A., Waters,J., Zhang,H. *et al.* (2016) Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. *Nat. Genet.*, **48**, 1119–1130.
86. Ross,E.M. and Markowetz,F. (2016) OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biol.*, **17**, 69.