Supplementary Material for
Modeling genome-wide replication kinetics reveals a mechanism for regulation of replication timing

Scott Cheng-Hsin Yang

Department of Physics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

E-mail:scotty@sfu.ca

Nicholas Rhind

Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, Worcester, MA 01605, United States

E-mail: nick.rhind@umassmed.edu

John Bechhoefer

Department of Physics, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Table of Contents

A. Limitations of the data

Although the microarray experiments analyzed here provide high-quality data, artifacts and limitations should be addressed. The brief description of the experiment below follows Alvino *et al* 2007 and Raghuraman *et al* 2001. The data from McCune *et al* 2008 were obtained using similar procedures.

Budding yeast cells were grown in an isotopically dense (^{13}C, ^{15}N) medium for a few generations at 23 °C and then synchronized at G1 by exposure to alpha mating pheromone. The culture was then resuspended in an isotopically light (^{12}C, ^{14}N) medium and further synchronized at the G1/S boundary by incubation at 37 °C, the restrictive temperature for *cdc7-1*. When 93% of the cells were budded, the temperature was lowered to the permissive temperature 23 °C to allow cells to enter S phase. Samples were collected throughout S phase. The DNA of the collected cells was first fragmented with a restriction enzyme (Eco RI). Dense and light DNA were then separated by ultracentrifugation, separately labeled with Cy3-dUTP and Cy5-dUTP, and hybridized to a open-reading-frame microarray. The intensities, after normalization by the mass of the sample, were used to calculate the fraction of replication (Alvino *et al*, 2007).

A limitation of the data is its resolution. The data covers roughly the entire genome at time points from 10 to 45 minutes, as measured from the release of the *cdc7-1* block. It comprises 8 time points (with 5-minute temporal resolution) and, on average, 6149 position points for each time point (spatial resolution = genome size / number of points ≈ 12000 kb / 6149 ≈ 2 kb). The average spatial resolution of 2 kb cannot resolve every single origin. In our fits, the elimination of origins that are less than 5 kb apart from their neighbors reflects this limitation. Given that the fork speed $v \approx 2$ kb/min, a typical origin can cover roughly 60 kb (average $t_{1/2} \times v$) of DNA. Thus, treating all origins in the region $x_i \pm 2.5$ kb as an effective origin at x_i would not change the replication fraction profiles. The average error on the x_i is 0.7 kb for the SM and 1.1 kb for the MIM.

The exact number of effective origins that we found depends on the elimination criteria (see Methods), as some origins made only marginal contributions to the replication profiles. The parameters of these origins have relatively large errors (t_w± 50% for the SM; n± 30% for the MIM). They were also sensitive to the form of data used in the fit (e.g., smoothed data versus raw; Supplementary Material Section F). The marginal origins constitute about

10% of all origins identified. Since they do not affect the replication program significantly (replication profile at 30 min changes by less than 5%), uncertainties about their numbers and parameters do not change the results presented above.

Another issue is that the data does not cover the entire range of possible replication fraction (0–100%); roughly all the data spreads between 10–90% (Supp. Fig. 1). One contribution to this artifact is the inability to cleanly separate the replicated fragments from the unreplicated. Alvino *et al* reported that small fragments and A-T rich sequences of unreplicated DNA are less dense and are physically similar to the replicated fragments. This leads to non-zero replication signals everywhere, even when no DNA is replicated. To understand the upper bound of 90%, we note that Alvino *et al* report that, for each time point, they normalized the microarray signals by the ratio between the total signal and the DNA fragments' total mass (Alvino *et al*, 2007). Although the normalization corrects for large amounts of signal drifts and scaling, we suspect that the rescaling is not perfect. To compensate for the reduced range of replication, we introduced a global background and a constant scaling factor for each time point as (genome-wide) parameters. Since these parameters are genome wide, they affect all origins simultaneously. However, the relationships between the local SM parameters $t_{1/2}$ and t_w are not significantly affected. Similarly, the relative values of the MIM local parameter n are also not significantly affected.

In the microarray experiment, the progress of replication is monitored with flow cytometry (Alvino *et al*, 2007). The flow-cytometry data shows that DNA content stopped increasing after 60±10 minutes into S phase (Alvino *et al*, 2007; Fig. 1A). We therefore estimate S phase to be 60 min. With our definition of potential efficiency ($\Phi(t_{end})$), a change in t_{end} changes the potential efficiency of every origin. (Potential efficiencies as a function of t_{end} can be estimated from Fig. 3B.) Still, the trend that later-firing origins have lower potential efficiency remains valid, as is the trend between observed and potential efficiency shown in Fig. 4C.

B. Statistical details of the fits

Here, we discuss in more detail the various fits described in the main text. We start by recalling the definition of the χ^2 statistic:

$$\chi^2 = \sum_{i=1}^{N_d} \frac{(f_i - d_i)^2}{\sigma_i^2}\,, \tag{1}$$

where N_d is the number of data points, f_i is the model value, d_i is the data (measurement) value, and σ_i is the standard deviation of the measurement d_i. Use of the χ^2 statistic (least-squares fitting) asserts that statistical fluctuations affect each data point d_i independently and that the fluctuations are Gaussian distributed, with mean 0 and standard deviation σ_i, which we denote $\mathcal{N}(0, \sigma_i)$. As we shall see, a detailed examination of the fluctuations shows that the assumptions for least-squares fits are not strictly met.

Ideally, the noise distribution for each data point would be estimated by repeating the experiment a large number of times. McCune *et al* repeated their experiment once, meaning that there are just two measurements of each data point. To examine the distribution of fluctuations, we considered the distribution of the differences between the experiments, calculated data point by data point. (Supp. Fig. 4A). A cursory examination shows that the fluctuations vary notably with time: earlier time points show smaller fluctuations than later ones. We thus grouped the fluctuations by time points. Within each time point, fluctuations are homogeneous, except for an obvious upward bias corresponding to the data points representing chromosome I (Supp. Fig. 4B). We observed a similar bias in all 8 time points and thus excluded the data from the set of residuals used to estimate the distribution of fluctuations. (However, we did not exclude chromosome I from our model fits.)

Excluding the differences from chromosome I, we compiled histograms for the 8 time points (Supp. Fig. 4C). These histograms estimate probability distribution functions for the differences between two noisy measurements. For curve fitting, we need to estimate the distribution of a *single* noisy measurement. Elementary properties of the variance imply that, for two independent random variables X and Y, $\mathrm{Var}[X - Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$. For two independently and identically distributed random variables, the standard deviations of the differences are then $\sqrt{2}$ times larger than the standard deviation for single-measurement noise. Correcting for this factor, we found the following standard deviations: $\sigma_{10} = 1.16$, $\sigma_{15} = 1.43$, $\sigma_{20} = 1.96$, $\sigma_{25} = 2.46$, $\sigma_{30} = 2.96$, $\sigma_{35} = 3.37$, $\sigma_{40} = 3.05$, and $\sigma_{45} = 3.00$ (The caption to Supp. Fig. 4C gives the uncorrected values.)

To examine the fluctuation distributions further, we rescaled the fluctuations for each data point by dividing by the standard deviation for that time point. After rescaling, all eight histograms collapse to a single distribution (Supp. Fig. 4D). This confirms that the noise fluctuations depend only on a reduced variable $(f_i - d_i)/\sigma_i$, as assumed when writing down Eq. 1. Unfortunately, two problems need to be addressed in order to perform a rigorous fit. First, the distributions are not Gaussian distributed (Supp. Fig. 4C). In particular, the positive-valued tails are approximately exponential, implying that large fluctuations are much more likely than a noise model (likelihood function) based on Gaussian statistics would suggest. Second, the distribution is clearly skewed (asymmetric about 0). This means that the noises of the two experiments are not identically distributed and that the σ values obtained from the $\sqrt{2}$ scaling might not approximate the deviation of the single-measurement noise well. (It is easy to prove that the difference between two independently and identically distributed random variables must be distributed symmetrically about zero.) Without more measurements, it is difficult to infer the actual form of the noise distribution. One further test examines the independence of fluctuations in one data point compared to another. We checked this by computing the autocorrelation function of the (scaled) residuals. The autocorrelation curves collapse, and there is only a weak correlation in the first few data points (Supp. Fig. 4E). Thus, the assumption of independence is reasonably well satisfied.

At this point, we have established that it is reasonable to treat the fluctuations in each data point separately and that the fluctuations are a function only of the reduced variable $(f_i - d_i)/\sigma_i$. Although we do not know the exact form of the likelihood function, we can examine how sensitive our model fits are to its precise form. Thus, we compared least-squares fits (assumes Gaussian likelihood function) and robust fits (assumes exponential deviations and uses a $\chi^2 = \sum_{i=1}^{N_d} |f_i - d_i|/\sigma_i$). The comparison used the data from chromosome XI, and some of the results are shown in Supp. Fig. 5. In general, we found little to distinguish between the results of the two fits. The main difference is that there are systematic shifts between corresponding parameters. The robust fit shifts the global v by ≈ -0.2 kb/min, origin positions by $\approx \pm 1$ kb, $t_{1/2}$ by ≈ -3 min, t_w by ≈ -2 min, and n by ≈ -1. We speculate that using the actual noise distribution to fit would give parameters whose values are inbetween those obtained from a least-squares fit and those obtained from a robust fit.

Since least-squares and robust fits give similar parameter values, we decided to adopt the more standard least-squares χ^2 statistic, noting however that any P values will be severely underestimated, as they fail to account for the exponential tail of the distribution. For a similar reason, the statistical errors for the parameters estimated by the fit will be underestimated. We listed them nonetheless, as their relative values attest to the relative certainty in the associated fit parameters of the same type.

In reporting our fits, we follow common practice and record, instead of χ^2, the "reduced chi square" $\chi^2_\nu \equiv \chi^2/\nu$, where ν is the number of degrees of freedom, $\nu = N_d - N_p$, with N_d the number of data points and N_p the number of free parameters in the fit. For $\nu \gg 1$, always true in our analysis, the χ^2_ν statistic is expected to be distributed as $\mathcal{N}(1, \sqrt{2/\nu})$. However, we recall that the exponential tail of the noise fluctuations will increase the expected standard deviation of the χ^2_ν statistic significantly.

Before proceeding to whole-genome fits, we first made a detailed comparison of the VVSM, SM, and MIM models on chromosome XI, which has $N_d = 2678$ and $N_p = 99$, 76, and 54 for the VVSM, SM, and MIM, respectively. The χ^2_ν values for the three models are 2.29, 2.48, and 2.76. These values exceed the expected χ^2_ν value of 1 by 42, 53, and 63 standard deviations. Given the uncertainty in the distribution of χ^2_ν, we did not reject the fits but attempted a more qualitative description of the fit quality (Supp. Fig. 6). The fit residuals and their distributions are all quite similar (Supp. Fig. 6A and B). The autocorrelation function is only slightly larger than that for the noise estimate (Supp. Fig. 6C), suggesting that the fits do capture most of the details of the data. The similarity of results for the three models justifies favoring the model with fewest parameters (MIM model). Repeating the comparison for whole-genome fits, we found χ^2_ν for the SM and MIM genome-wide fits: 4.91 and 5.83 ($\nu = 48129$ and 48481).

C. Comparison between models with variable and constant fork velocity

The formalism introduced in the Methods can be extended to incorporate a space-time-dependent fork velocity $v(x,t)$. We generated a spatially varying $v(x)$ as follows: The summand in Eq. 7 in the main text is only non-zero when Δx_p contains an origin at x_i, implying that the sum is really over $p = i$. By replacing the global v by a local v_i, we associated a different fork velocity with each origin. In this way, we obtained spatially varying fork velocities. Generalizing further, with a variable fork velocity $v(t)$, the edges of the triangle in Fig. 7 would be curved. The goal is then to find the time along the curved edge by solving

$$\int_{t_e}^{t} v(t)dt = |x - x_p| \tag{2}$$

for t_e. Here, t_e is a function of t, $|x - x_p|$ and the parameters that form $v(t)$. This generalizes the constant-velocity case, where $t_e = t - |x - x_p|/v$. Replacing the argument $t - |x - x_p|/v$ used previously with $t_e(t, |x - x_p|, v_{i,\cdots})$ [with $v_{i,\cdots}$ representing the parameters that describe $v(t)$], one obtains a formalism that allows for a time-dependent fork velocity. In the fits, we kept the velocity constant in time. This is consistent with independent evidence that the velocity is constant throughout S phase (Rivin & Fangman, 1980).

We used this "variable-velocity-sigmoid model" (VVSM), the SM, and the MIM to fit chromosome XI (Supp. Figs. 7). Each of the three models captures most of the variations in the data, explaining 98.87% (VVSM), 98.77% (SM), and 98.62% (MIM) of the variance of the raw data. Below, we also showed that the distribution of the residuals of the three fits are very similar (Supp. Fig. 6B), indicating that the goodness of the three fits are similar. Thus, we conclude that constant-velocity models describe the replication kinetics as well as variable-velocity models.

D. Mean-field analysis of origin efficiency

The relationship between efficiency and potential efficiency shown in Fig. 4 can be mostly explained by a mean-field analysis. The idea is that all the neighboring origins of an origin are replaced by an "average neighbor" whose firing-time distribution is the average of all the distributions. We averaged over all 342 firing-time distributions in the SM to produce the genome-wide-averaged $\phi_{avg}(t)$. We then computed the average nearest-neighbor distance ($\approx$ 28 kb) to locate the average neighbor. Next, we approximated t_w as a function of $t_{1/2}$ by fitting a power-law through Fig. 3D. The analytic relationship between t_w and $t_{1/2}$ implies that the potential efficiency is also a smooth function of $t_{1/2}$. Finally, the efficiency was then calculated by placing the average neighbor at the average nearest-neighbor distance beside origins. Going through all the $t_{1/2}$ values extracted, we generated the curve shown in Fig. 4C. This analysis suggests that the geometric effect we see on observed origin efficiency is not specific to the particular arrangement of origins in budding yeast; however, such an effect would be generally expected for a genome with this density of origins.

E. Effects of asynchrony in cell population

It is apparent that asynchrony widens firing-time distributions. Consider a scenario where the timing of every origin is deterministic. Since cells in an asynchronous culture enter S phase at different times, the initiation times would appear to be stochastic. To assess the effect of asynchrony on the parameters we extracted, we extended our formalism to include asynchrony.

For the modeling, we first distinguish between "starting-time asynchrony" and "progressive asynchrony." For the microarray experiment analyzed, the cell culture was synchronized in two steps (first by alpha-factor incubation then with *cdc7-1* block) before samples were taken for hybridization. We define starting-time asynchrony as the asynchrony of release from the last synchronization procedure. In other words, this is the asynchrony inherent to the synchronization methods used. Now, consider a scenario where the synchronization procedures produce a perfectly synchronized cell culture. If the replication program is not strictly deterministic, the DNA content for each cell would evolve differently as S phase proceeds. This "progressive asynchrony" is inherent to the stochastic replication program. The probabilistic model presented in the Methods captures precisely the effects of progressive asynchrony on microarray replication fraction profile. Since the data analyzed contains both types of asynchrony, we extend the formalism to include starting-time asynchrony.

We model the starting-time asynchrony of a cell population by a starting-time distribution $\psi_a(t)$, defined as the number density of the cell population that is t minutes into S phase. Cells associated with negative t enter S phase t minutes after the start of S phase. If the probed cell culture has a starting-time distribution $\psi_a(t)$, the measured replication fraction profile (containing both types of asynchrony) is expressed as the convolution

$$f_a(x,t) = \int_{-\infty}^{\infty} f(x,t')\psi_a(t-t')dt', \tag{3}$$

with $f(x,t)$ being the replication fraction profile for a cell culture having no starting-time asynchrony $[\psi_a(t) = \delta(t)]$.

We simulated the replication fraction profiles that contain both types of asynchrony using a slightly modified version of our previously developed method (Jun *et al.*, 2005). The theoretical prediction matches the simulation data well (Supp. Fig. 8A). The most apparent effects of the starting-time asynchrony are the "squeezing" of the peaks in the replication-fraction axis and the "stretching" of the peaks in the position axis. The former mainly translates into a wider firing-time distribution, whereas the latter translates into a faster fork progression rate.

To apply Eq. 3 to our analysis, we need an estimate of the starting-time distribution. To our knowledge, although there are works that estimate the starting-time distribution resulting from alpha-factor synchronization (Niemistö *et al*, 2007; Orlando *et al*, 2007), there are none related to the *cdc7-1* block. Since the *cdc7-1* block is the final synchronization step taken and since it blocks cells at the G1/S boundary, it is important to use an estimate of $\psi_a(t)$ that includes the effects of *cdc7-1.* To do this, we compared the flow-cytometric determination of DNA content between the 0-min and 20-min time points (Avino *et al.* 2007; Fig. 1A).

We measured the width by measuring the spread of DNA content at half the peak height. The width at 0-min is a reference width corresponding to perfect synchrony, as all the cells have 1C amount of DNA. The width at 20-min includes both types of asynchrony and can be used to generate an upper bound of the starting-time asynchrony. A simple image analysis shows that the full width of the 20-min peak is 5 pixels larger than that of the 0-min peak. DNA content increases from 1C to 2C over 75 pixels. Using a crude estimate that DNA content linearly increases with progression of S phase, we converted 5 pixels to 4 minutes (via 60 min/75 pixel). Since the flow-cytometric peaks are Gaussian-like, we set $\psi_a(t)$ to a normal distribution with mean zero and standard deviation 2 min, denoted by $\mathcal{N}(0, 2)$. The estimated asynchrony implies that 95% of the cells would have entered S phase within 8 min of the start of S phase.

We refit the SM to the chromosome-XI part of the data with Eq. 3 (instead of Eq. 7 in main text) and the estimated asynchrony. We found that the local parameters extracted with asynchrony are not significantly different from those extracted without (Supp. Fig. 8B and C). The estimated starting-time asynchrony shifts $t_{1/2}$ by $\approx$ -0.5 min, t_w by $\approx$ -1 min, and v by $\approx$ -0.3 kb/min. These shifts do not change the relationship between $t_{1/2}$ and t_w, and the results presented in the main text remain valid. We note that using a linear relationship between DNA content and time underestimates the asynchrony; however, refitting using $\mathcal{N}(0, 4)$ also gives shifts that are unimportant.

F. Fits to raw and smoothed data

It is common practice to analyze a smoothed version of microarray data so that peaks can be more easily identified. It is thus tempting to use smoothed data for curve fitting, as well. However, there are reasons to prefer fits to the raw, unsmoothed data. First, as a matter of principle, smoothing can only reduce the information available in a dataset and can never add to it. Second, the smoothing procedure correlates the statistical fluctuations among nearby data points, requiring a modification of standard least-squares fitting algorithms.

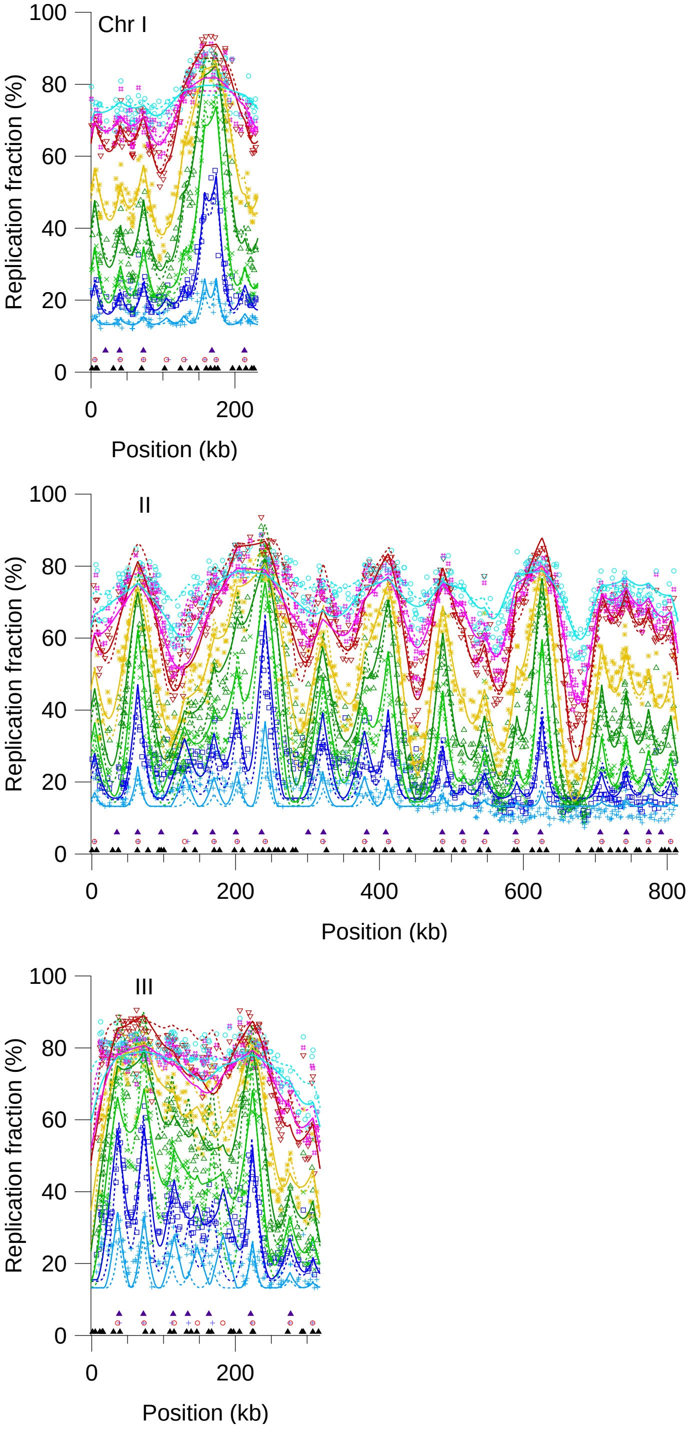
To test whether there are significant differences between the results of fitting to raw and to smoothed datasets, we repeated the SM fit of Chromosome XI using the smoothed data of McCune *et al* 2008. The residuals (Supp. Fig. 9A) and their autocorrelation function (Supp. Fig. 9B) show a correlation among neighboring points that results from the smoothing operation. The χ^2 statistic of standard least-squares routines then needs to be modified to explicitly account for the correlations (Sivia & Skilling, 2006), and using the standard statistic (Eq. 1) can bias the resulting parameter values. With this particular dataset, we found little practical difference between fitting to the raw data and fitting to the smoothed data using the standard χ^2 statistic. Both fits produced parameter values that mostly agreed to within 10%; only a few parameters, corresponding to less-apparent peaks in the microarray data, do not match well (Supp. Fig. 9C). Thus, it is unlikely that any substantive conclusions reached about this particular dataset would have been affected had we fit to the smoothed data, using the standard χ^2 statistic; however, since it is just as easy to fit raw data as it is to fit smoothed data, we recommend doing the former and encourage experimental groups to publish and make available the raw datasets.

Supplementary Table Legends

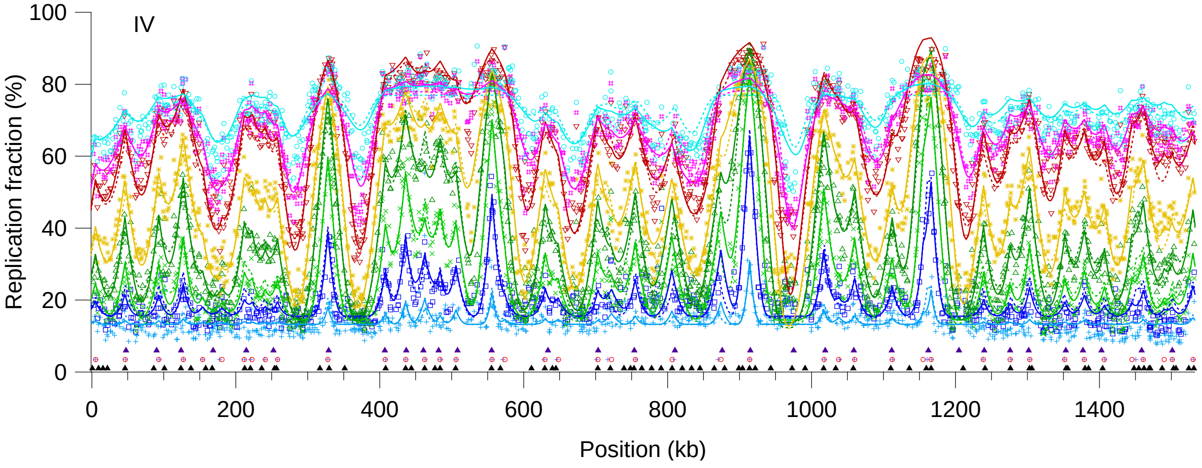
SUPP. TAB. I. Origin properties extracted from the genome-wide SM. For the column titles, we used the following abbreviation: “chr” for chromosome, “ori pos” for origin position, “err” for error, “pot eff” for potential efficiency, and “obs eff” for observed efficiency. Under the “Alvino,” “OriDB,” and “MIM” columns, 1 denotes that the origin is also identified in Alvino *et al* 2007, Nieduszynski *et al* 2007, and in the MIM, respectively, while 0 denotes not identified.

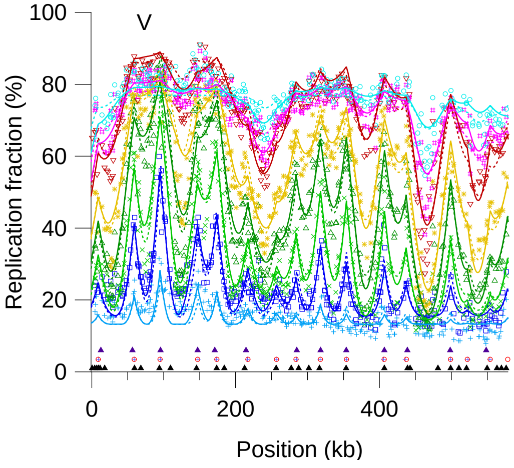
SUPP. TAB. II. Origin properties extracted from the genome-wide MIM. Same convention as Supp. Tab. I.

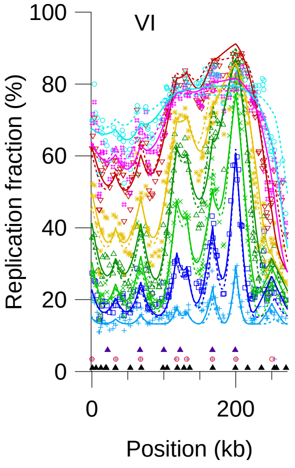
SUPP. TAB. III. Genome-wide parameters extracted from the SM and MIM fits. For the MIM, $t^*_{1/2}$ and r^* are used to construct the global $\phi_o(t) = t/[t + (t^*_{1/2})^{r^*}]$ (see Methods). The quantity $t^*_{1/2}$ plays a role that is analogous to the quantity $t_{1/2}$ for the SM model.

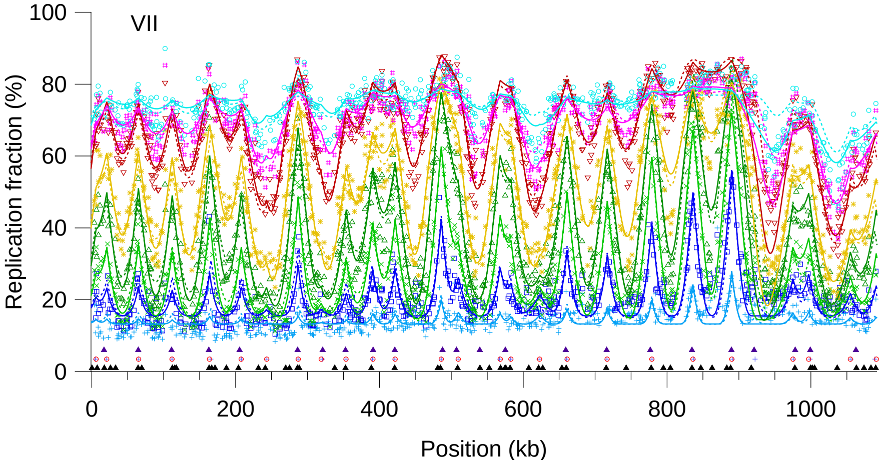


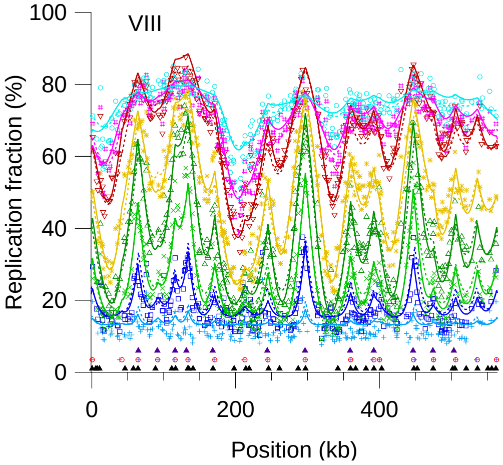
Supplementary Figures

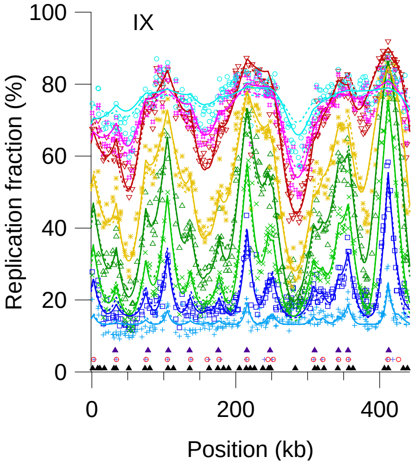

IV
Replication fraction (%)
100
80
60
40
20
0
0
200
400
600
800
1000
1200
1400
Position (kb)

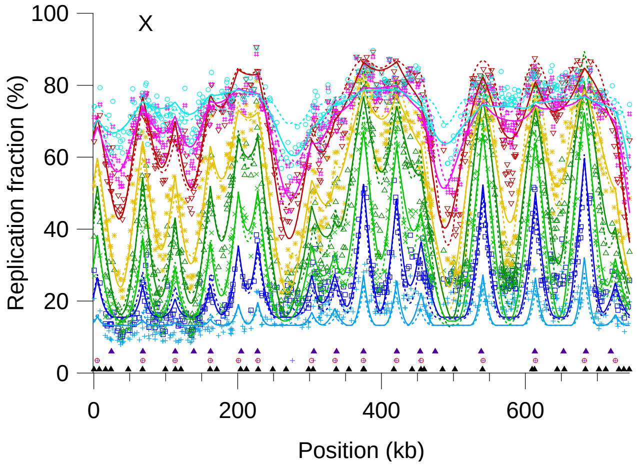

V
Replication fraction (%)
100
80
60
40
20
0
0
200
400
Position (kb)

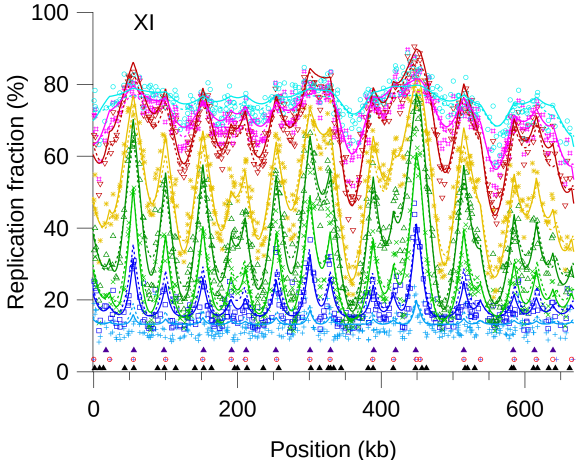

VI
Replication fraction (%)
100
80
60
40
20
0
0
200
Position (kb)

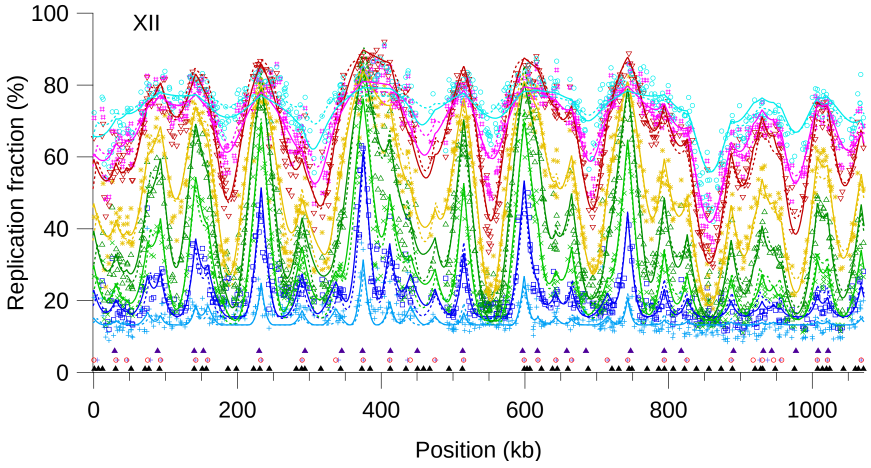

VII
Replication fraction (%)
100
80
60
40
20
0
0
200
400
600
800
1000
Position (kb)

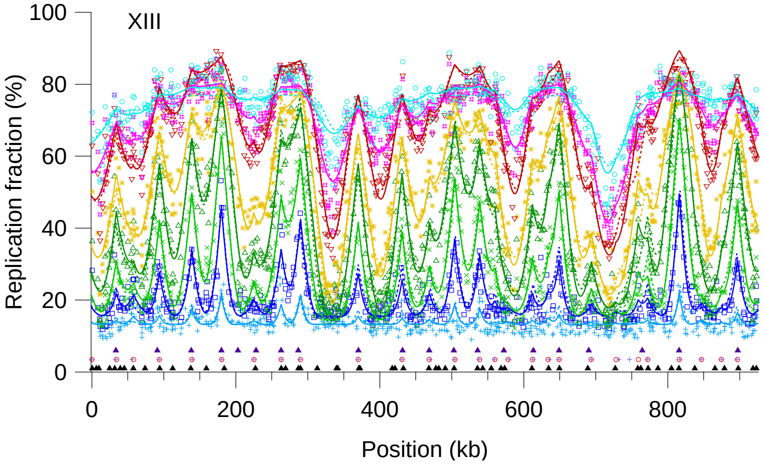

VIII
Replication fraction (%)
100
80
60
40
20
0
0
200
400
Position (kb)

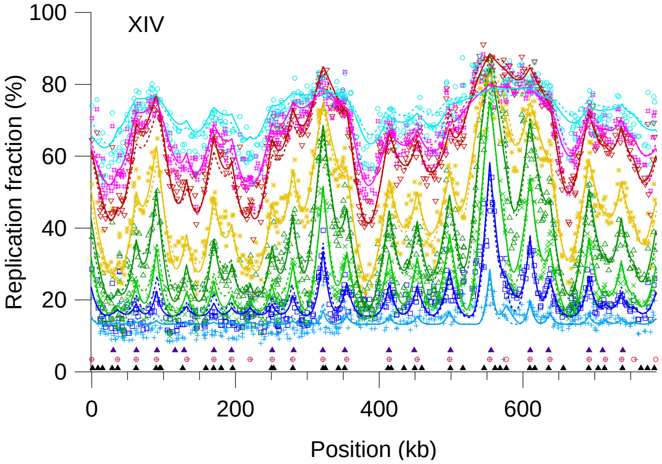

IX
Replication fraction (%)
100
80
60
40
20
0
0
200
400
Position (kb)

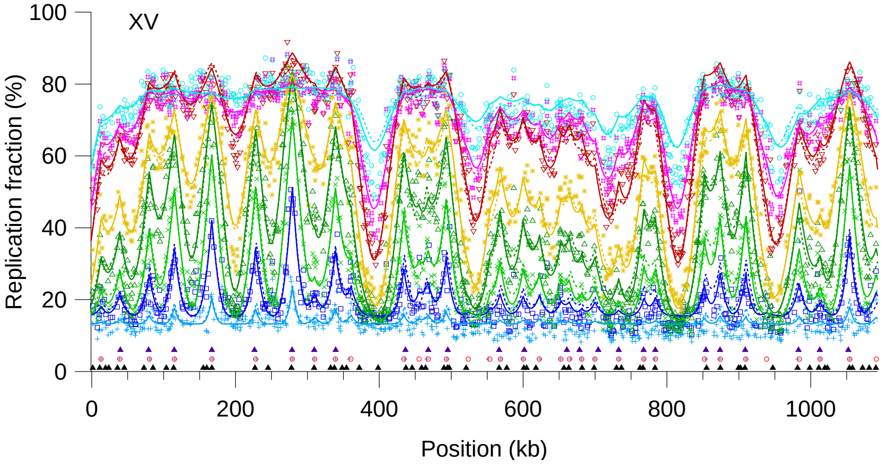

X
Replication fraction (%)
100
80
60
40
20
0
0
200
400
600
Position (kb)

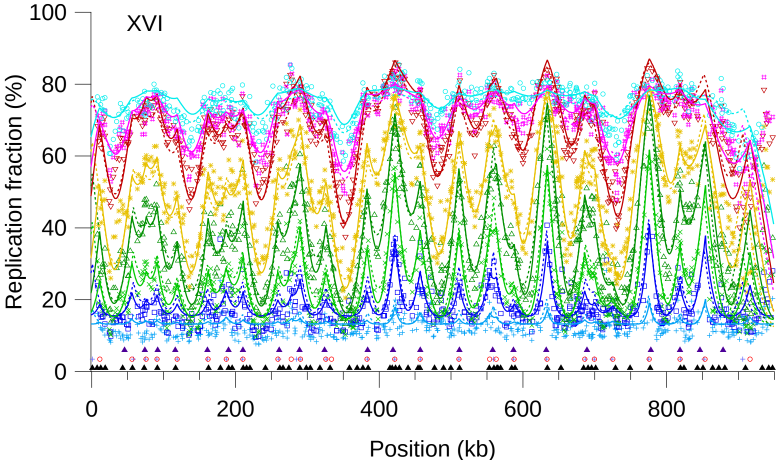

XI
Replication fraction (%)
100
80
60
40
20
0
0
200
400
600
Position (kb)


XII
Replication fraction (%)
100
80
60
40
20
0
0
200
400
600
800
1000
Position (kb)

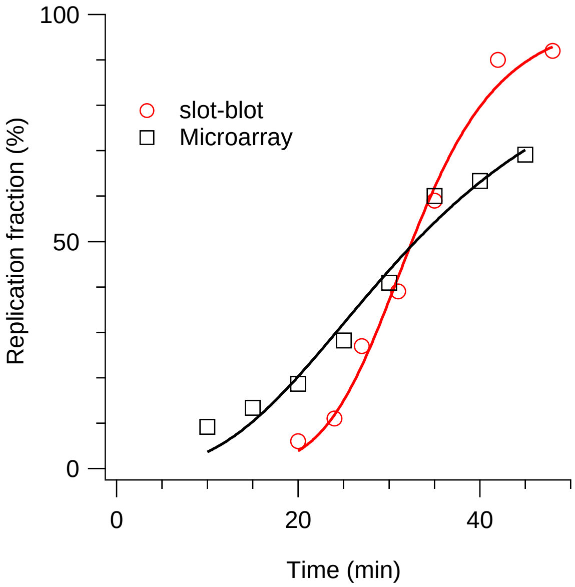

XIII
Replication fraction (%)
100
80
60
40
20
0
0
200
400
600
800
Position (kb)


XIV
Replication fraction (%)
100
80
60
40
20
0
0
200
400
600
Position (kb)

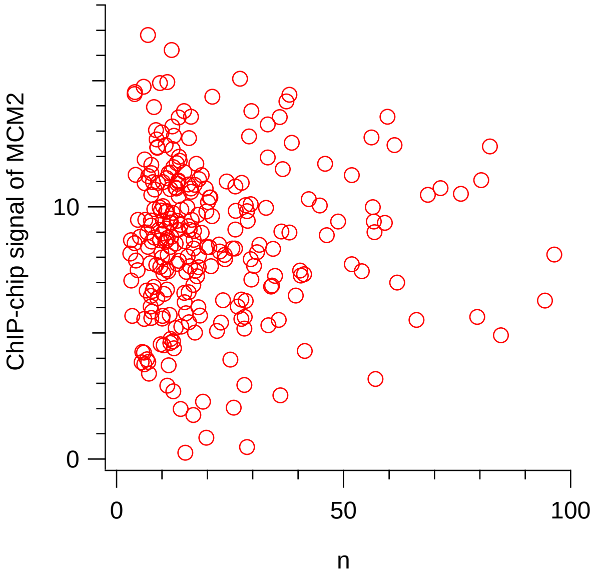

XV
Replication fraction (%)
100
80
60
40
20
0
0
200
400
600
800
1000
Position (kb)



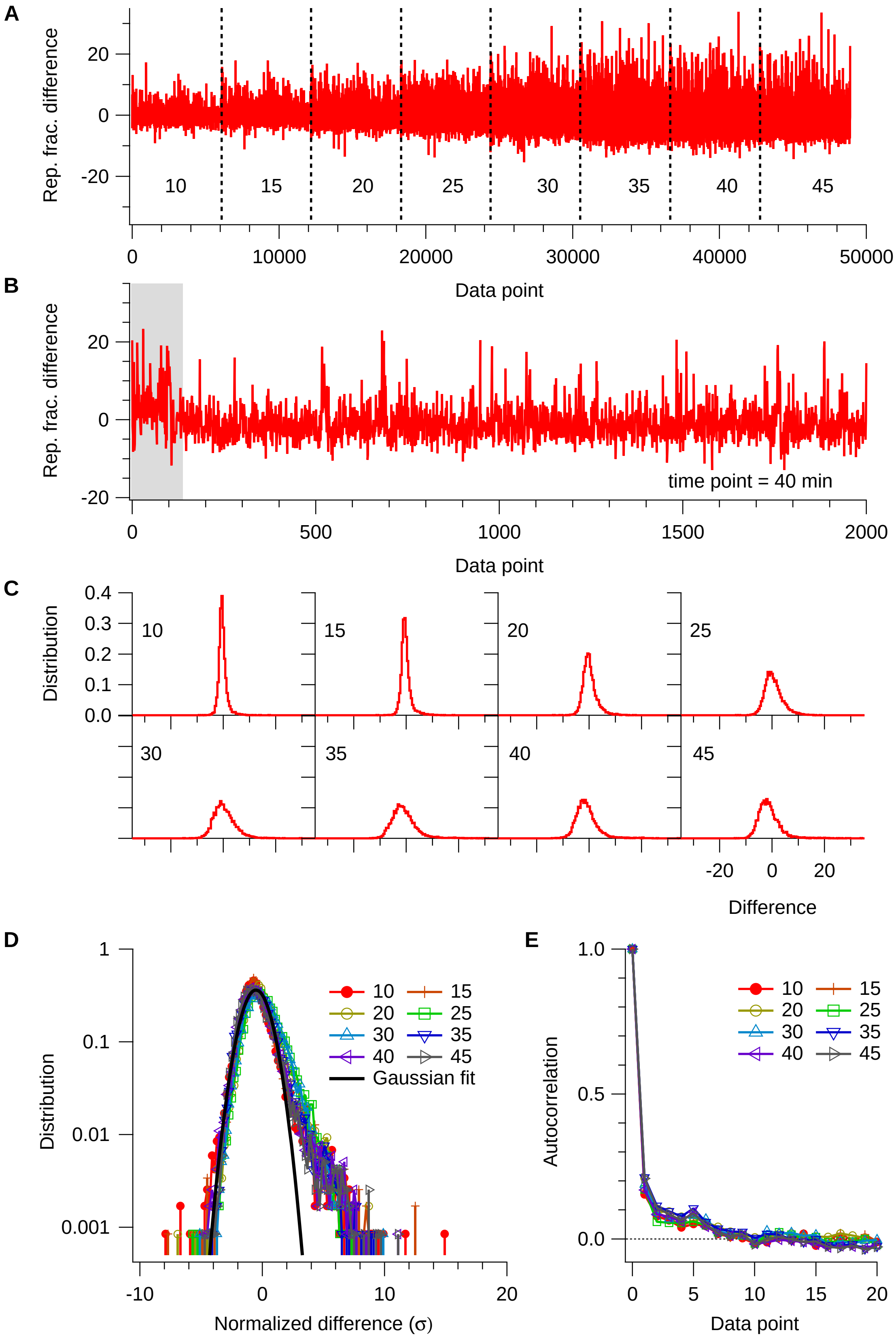
SUPP. FIG. 1: Genome-wide SM and MIM fits, separately shown for each chromosome. Roman numeral corresponds to chromosome number. The x-axis denotes the position along the chromosome. Markers are data; solid lines are fits from SM; dotted lines are fits from MIM. Upper row of solid triangles at the bottom denote origin positions identified in Alvino *et al* 2007. In the middle row, open circles correspond to estimated origin positions from the SM, while crosses correspond to those from the MIM. The lower row of triangles correspond to origins in the OriDB database (Nieduszynski *et al*, 2007). The eight curves from bottom to top correspond to the replication fraction $f(x)$ at 10, 15, 20, 25, 30, 35, 40, and 45 min after release from the *cdc7-1* block.



SUPP. FIG. 2: Replication fraction of ARS501. ARS501 is located on chromosome V at ≈ 549 kb. Circles are data from a slot-blot experiment (Ferguson *et al*, 1991); squares are data from the newer microarray experiment (McCune *et al*, 2008). Lines are fits to the data using a sigmoid (Hill equation). Values for t_{rep} and t_{width} are extracted for comparison. For the slot-blot, $t_{rep} = 33$ min and $t_{width} = 11$ min. For the microarray, $t_{rep} = 33$ min and $t_{width} = 26$ min.

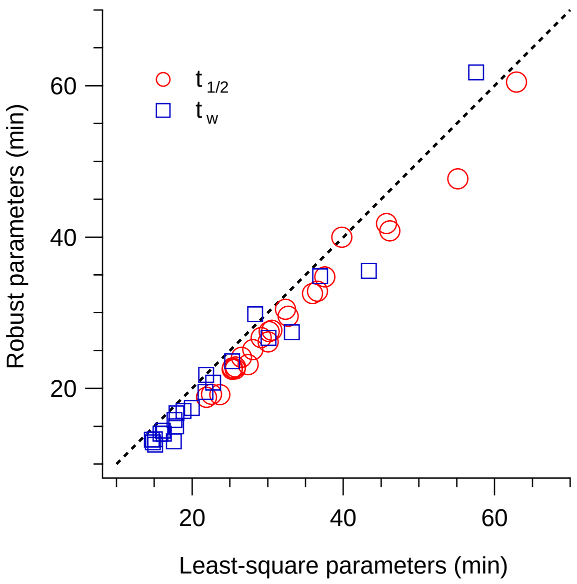


SUPP. FIG. 3: ChIP-chip signal vs parameter n. The y-axis is the ChIP-chip signal for MCM2 occupancy (Xu *et al*, 2006); the x-axis is the extracted parameter n from the MIM. Origins with larger n values are more efficient in the mode. The correlation coefficient between the two quantities is 0.003 which is less than the critical value indicating a correlation ($r_c = 0.121$, two-sided test, 264 degrees of freedom, signficance level = 0.05).

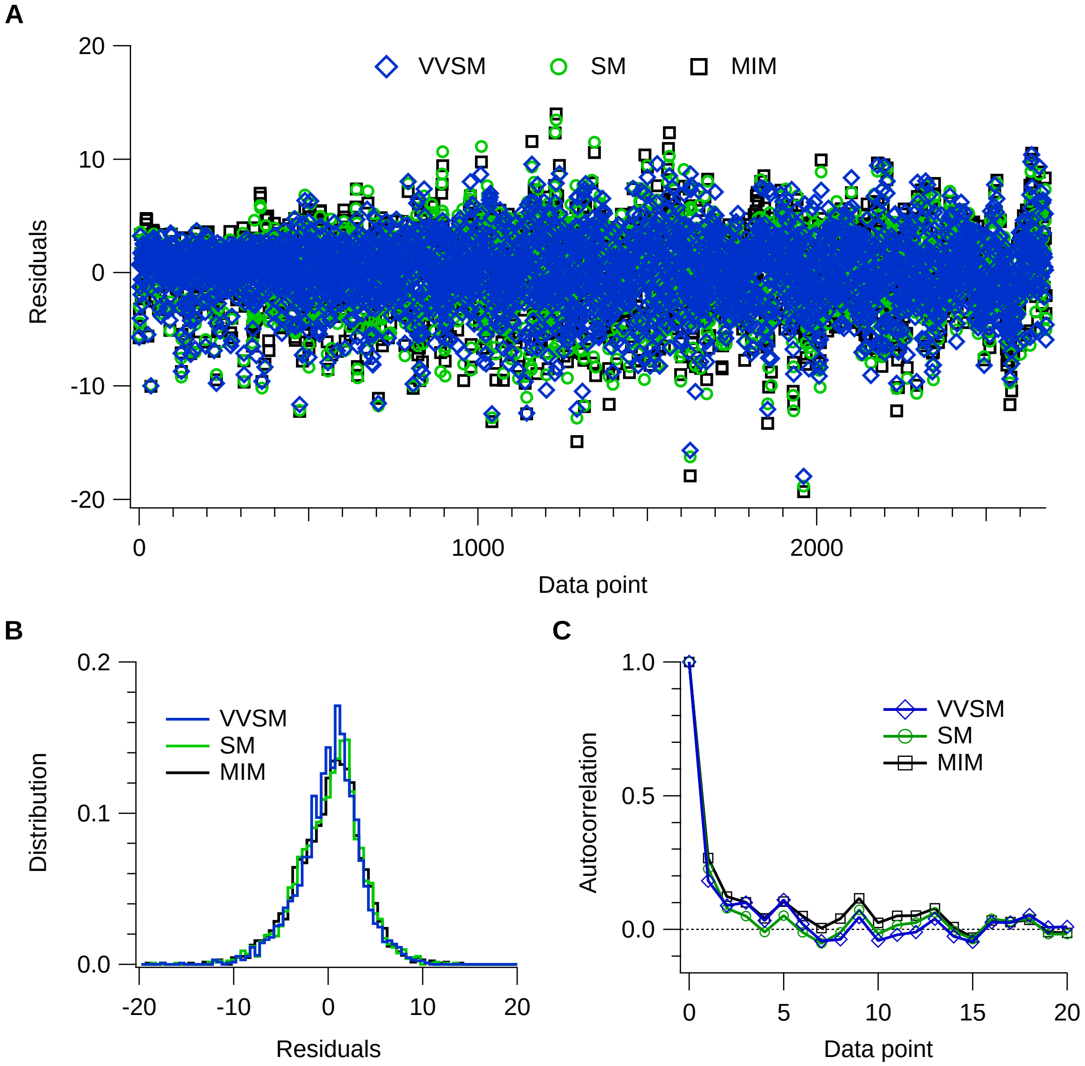

A
Rep. frac. difference
20
0
-20
10
15
20
25
30
35
40
45
0
10000
20000
30000
40000
50000
Data point
B
Rep. frac. difference
20
0
-20
time point = 40 min
0
500
1000
1500
2000
Data point
C
Distribution
0.4
0.3
0.2
0.1
0.0
10
15
20
25
30
35
40
45
-20
0
20
Difference
D
Distribution
1
0.1
0.01
0.001
10
15
20
25
30
35
40
45
Gaussian fit
-10
0
10
20
Normalized difference (σ)
E
Autocorrelation
1.0
0.5
0.0
10
15
20
25
30
35
40
45
0
5
10
15
20
Data point

SUPP. FIG. 4:

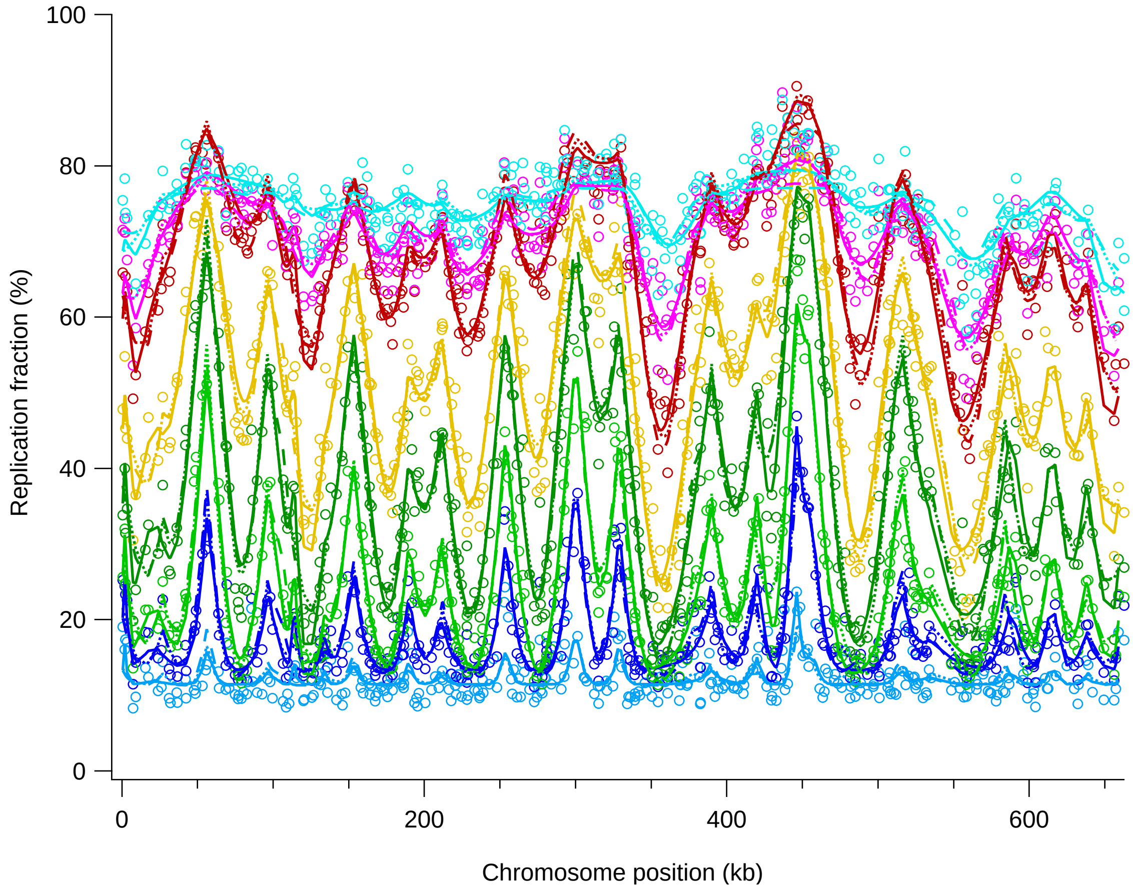
SUPP. FIG. 4: **A**. Difference between two equivalent experiments from McCune *et al* 2008. The differences between the replication fraction of two nominally equivalent experiment are shown serially in time. The fluctuation of the differences varies across different time points. **B**. Differences for the first 2000 data points time point 40 are shown. The upward bias in the shaded region corresponds to chromosome I. All time points have this bias. **C**. Histograms of the differences for different time points. In making the the histograms (bin width = 0.5), we excluded the first 200 data points of each time point because of the apparent upward bias. These data points correspond almost exactly to chromosome I. The standard deviation of the differences (in sequence of increasing time points) are 1.64, 2.03, 2.77, 3.48, 4.18. 4.76. 4.31, and 4.25. The σ_t values used in the fits equal these standard deviations divided by $\sqrt{2}$. **D**. The histograms (bin width = 0.2) in C collapse onto the same distribution after scaling the differences for each time point with its corresponding σ_t. Small deviations are Gaussian like, while large positive deviations are exponential. **E**. Autocorrelation of the differences. The autocorrelation shown excludes the first 200 data points, as well.



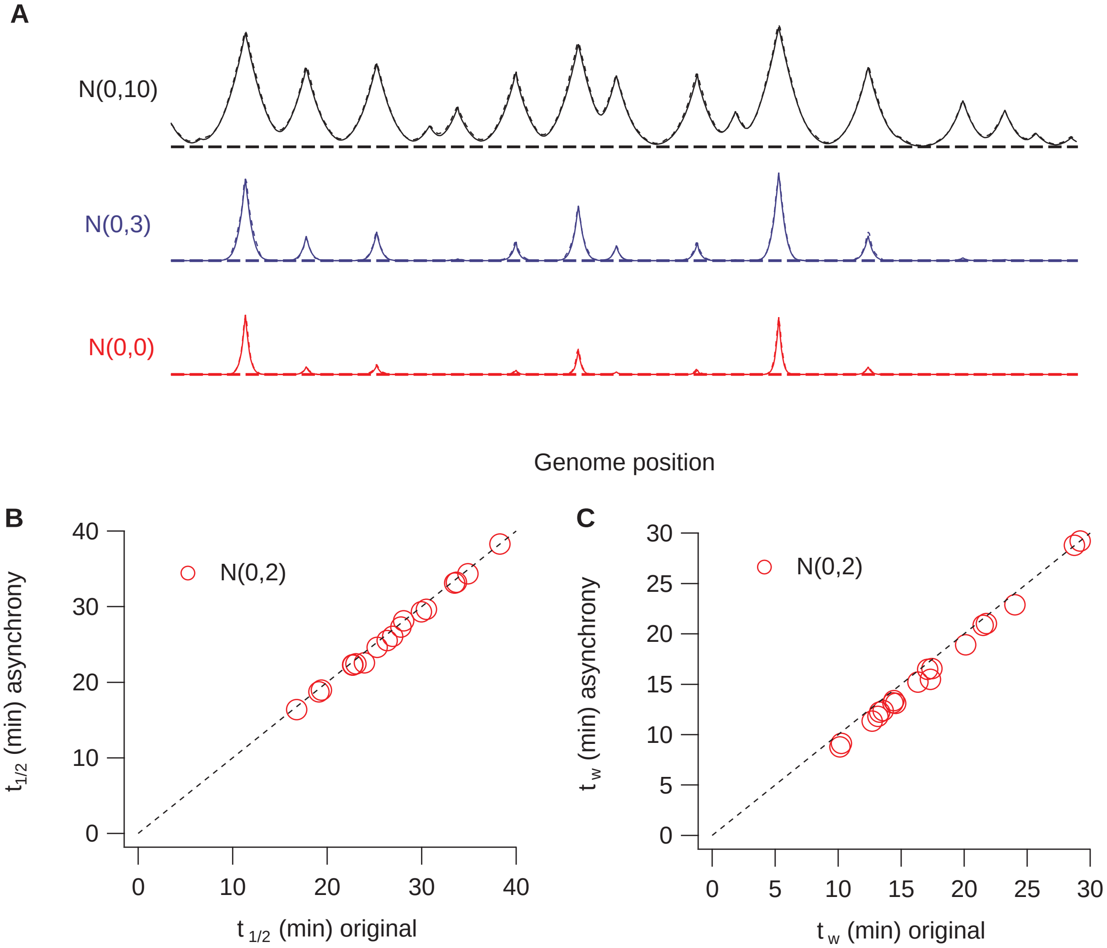
SUPP. FIG. 5: Comparison between least-squares and robust fit parameters for chromosome XI. The x-axis corresponds to the least-squares fit, the y-axis to the robust. Dotted line shows $y = x$. The least-squares $t_{1/2}$ (t_w) values are on average 3.24 (0.73) min larger than the robust $t_{1/2}$ (t_w) values.



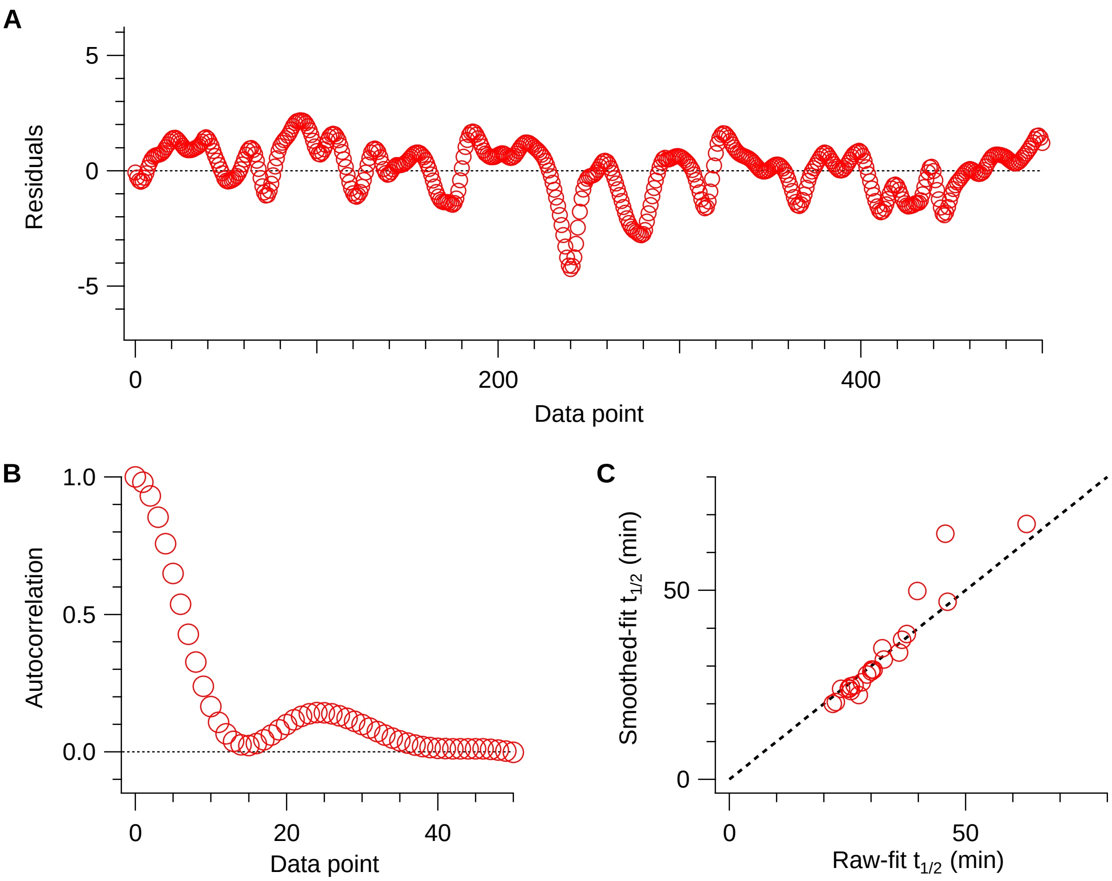
SUPP. FIG. 6: **A**. Residuals of the model fits to chromosome XI. Markers correspond to the residuals of the three different model fits, VVSM, SM, and MIM, discussed in the text. The residuals are plotted serially in time points. **B**. Histogram of the residuals with bin width = 0.5. The standard deviations of the VVSM, SM, and MIM residuals are 3.21, 3.42, and 3.52, respectively **c**. Autocorrelation of the residuals of the VVSM, SM, and MIM fits.



SUPP. FIG. 7: **A**. Fits to chromosome XI. Markers are data; solid lines are fits from VVSM; dotted lines are fits from SM; and dashed lines are fits from MIM. The eight curves from bottom to top correspond to the replication fraction $f(x)$ at 10, 15, 20, 25, 30, 35, 40 and 45 min after release from the restriction temperature of *cdc*7-1. The dataset covers the genome at ≈ 2-kb resolution.



SUPP. FIG. 8: **A**. Simulation and theoretical replication fraction profile with three different starting-time distributions. The notation N(μ,σ) denotes a normal distribution with mean μ and standard deviation σ. The three curves are generated using the same set of SM parameters (x_i, $t_{1/2}$, t_w and v) and correspond to the same time point. The only difference among them is the starting-time distribution. The theoretical calculation (solid cruves) matches the simulations (dashed curves) well. Horizontal dashed lines are the replication fraction 0-lines for the three cases. **B**. Comparison of $t_{1/2}$ fit parameters. The x-axis corresponds to the SM parameters extracted without consideration of asynchrony; the y-axis corresponds to the case with consideration of asynchrony. Dashed line shows $y = x$. **C**. Comparison of t_w fit parameters. The x-axis, y-axis, and dotted lines are as described in B.



SUPP. FIG. 9: **A**. Residuals of SM fit to the smoothed data of chromosome XI. the first 500 of the 5136 data points of residuals are shown for clarity. The number of data points here is larger than that of the raw data (2678) because the smoothed data was also interpolated (McCune *et al*, 2008; Raghuraman *et al*, 2001). **B**. Autocorrelation of residuals, showing the correlation in noise produced by the smoothing algorithm. **C**. Comparison of $t_{1/2}$ fit parameters. The x-axis corresponds to the raw data, the y-axis to the smoothed data. Dotted line shows $y = x$.

References

Alvino GM, Collingwood D, Murphy JM, Delrow J, Brewer BJ, Raghuraman MK (2007) Replication in hydroxyurea: it's a matter of time. *Mol Cell Biol* **27**: 6396–6404

Jun S, Zhang H, Bechhoefer J (2005) Nucleation and growth in one dimension. I. The generalized Kolmogorov-Johnson-Mehl-Avrami model. *Phys Rev E* **71**: 011908

McCune HJ, Danielson LS, Alvino GM, Collingwood D, Delrow JJ, Fangman WL, Brewer BJ, Raghuraman MK (2008) The temporal program of chromosome replication: genomewide replication in $clb_5\Delta$ *Saccharomyces cerevisiae*. *Genetics* **180**: 1833–1847

Nieduszynski CA, Hiraga S, Ak P, Benham1 CJ, Donaldson AD (2007) OriDB: a DNA replication origin database. *Nucleic Acids Res* **35**: D40–D46. `http://www.oridb.org`

Niemistö A, Nykter M, Aho T, Jalovaara H, Marjanen K, Ahdesmäki M, Ruusuvuori P, Tianinen M, Linne ML, Yli-Harja O (2007) Computational methods for estimation of cell cycle phase distribution of yeast cells. *EURASIP J Bioinfor Sys Biol* **2007**: 46150.

Orlando DA, Lin CY, Bernard A, Iversen ES, Hartemink AJ, Haase SB (2007) A probabilistic model for cell cycle distributions in synchrony experiments. *Cell Cycle* **6**: 478–488.

Rivin CJ, Fangman WL (1980) Replication fork rate and origin activation during the S phase of *Saccharomyces cerevisiae*. *J Cell Biol* **85**: 108–115

Sivia DS, Skilling J (2006) *Data Analysis: A Bayesian Tutorial.* Oxford University Press, New York, NY, USA