

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

Supplementary Material Section A: Limitations of the data

Supplementary Material Section B: Statistical details of the fits

Supplementary Material Section C: Comparison between models with variable and constant fork velocity

Supplementary Material Section D: Mean-field analysis of origin efficiency

Supplementary Material Section E: Effects of asynchrony in cell population

Supplementary Material Section F: Fits to raw and smoothed data

Supplementary Table Legends (see tables in spreadsheet)

Supplementary Figures 1-9

References

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).

B. Statistical details of the ts

caption to Supp. Fig. 4C gives the uncorrected values.)

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" $\chi^2_{\text{red}} = \chi^2 / \nu$, 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_{red} 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_{red} 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_{red} values for the three models are 2.29, 2.48, and 2.76. These values exceed the expected χ^2_{red} value of 1 by 42, 53, and 63 standard deviations. Given the uncertainty in the distribution of χ^2_{red} , 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_{red} 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 x_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

$$\int_{t_e}^t v(t) dt = jx - x_{pj} \quad (2)$$

for t_e . Here, t_e is a function of t , $jx - x_{pj}$ and the parameters that form $v(t)$. This generalizes the constant-velocity case, where $t_e = t - jx - x_{pj}/v$. Replacing the argument $t - jx - x_{pj}/v$ used previously with $t_e(t; jx - x_{pj}; v_i)$ [with v_i 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 ring-time distribution is the average of all the distributions. We averaged over all 342 ring-time distributions in the SM to produce the genome-wide-averaged $t_{avg}(t)$. We then computed the average nearest-neighbor distance (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 ring-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 $a(t)$, 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 $a(t)$, the measured replication fraction profile (containing both types of asynchrony) is expressed as the convolution

$$f_a(x; t) = \int_{-1}^1 f(x; t') a(t - t') dt';$$

translates into a wider ring-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

F. Fits to raw and smoothed data

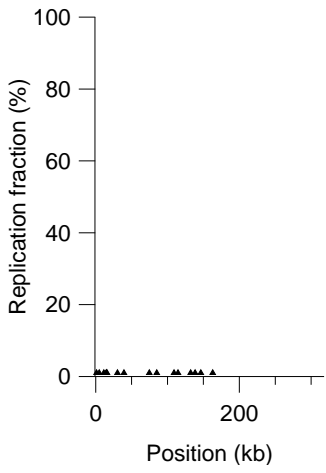
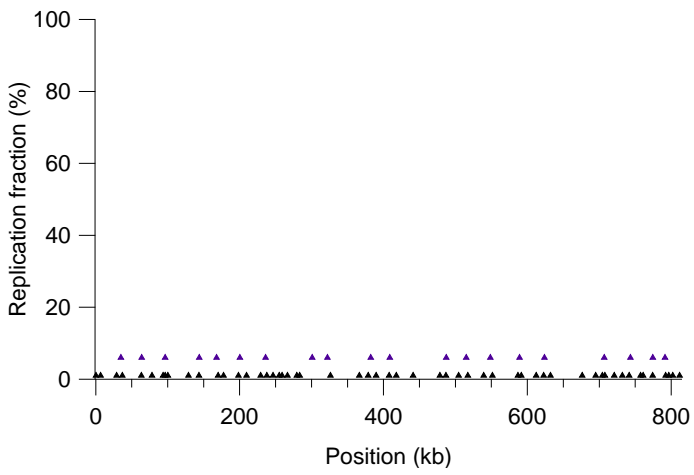
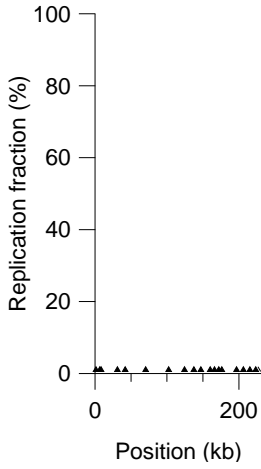
It is common practice to analyze a smoothed version of microarray data so that peaks can be more easily identified. It is thus tempting to use smoothed data for curve fitting, as well.

Supplementary Table Legends

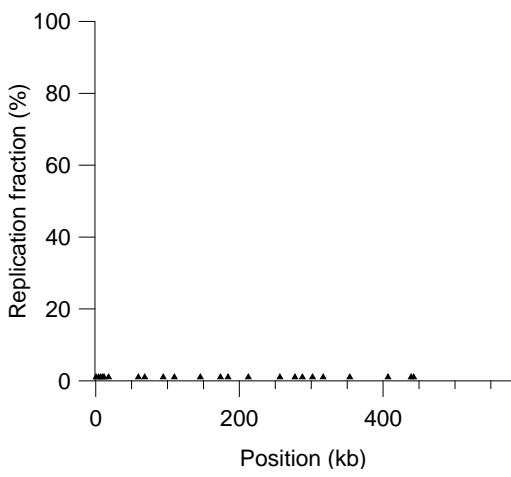
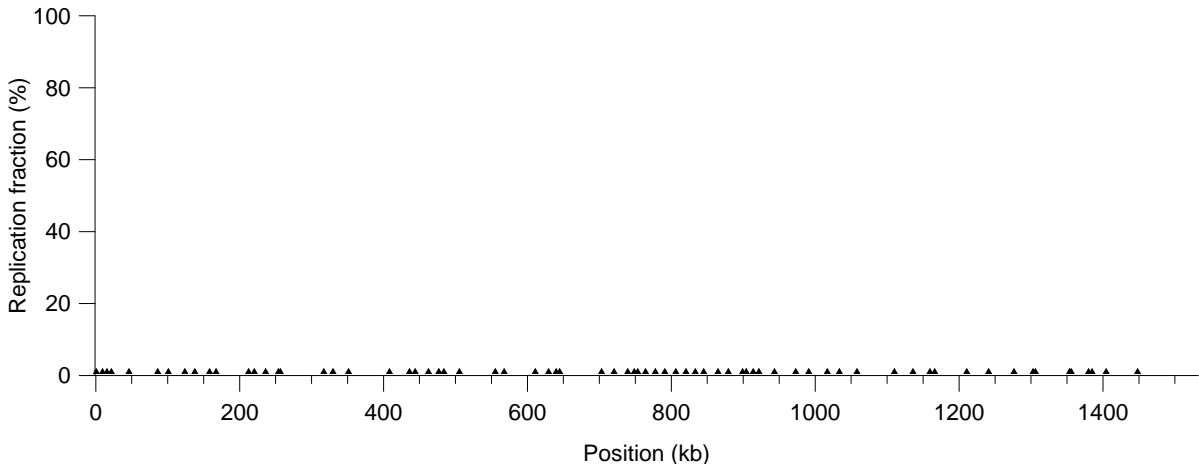
SUPP. TAB. I. Origin properties extracted from the genome-wide SM. For the column titles, we used the following abbreviation: \chr" for chromosome, \ori pos" for origin position, \err" for error, \pot e " for potential efficiency, and \obs e " for observed efficiency. Under the \Alvino,"theusOriDB,"theandusMIM"thecolumns,that origin is identified Alvino *et al* 2007,Nieduszynski *et al* 2007,andinectively,while not identified.

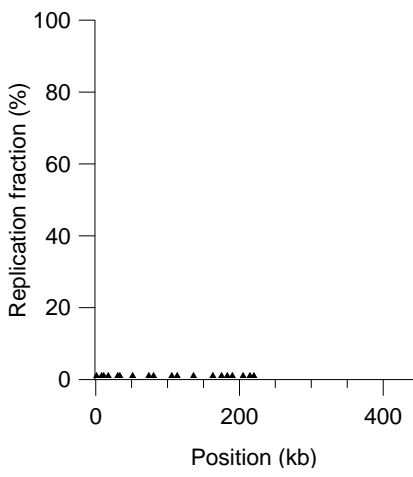
SUPP. TAB. II. Origin properties extracted from the genome-wide MIM. Same convention as Supp. Tab. I.

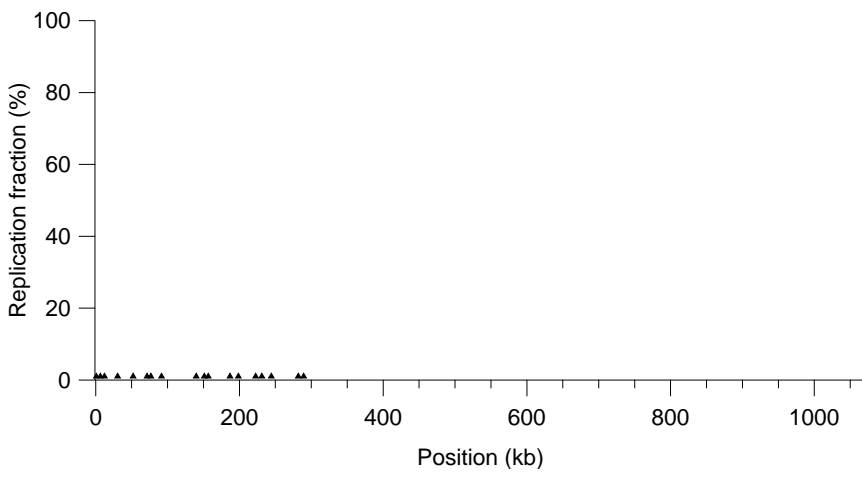
SUPP. TAB. III. Genome-wide parameters extracted from and For MIM, $t_{1=2}$ and r are used to construct the global $\rho(t) = t/[t + (t_{1=2})^r]$ (see Methods). The quantity $t_{1=2}$ plays a role that is analogous to the quantity $t_{1=2}$ for the model.

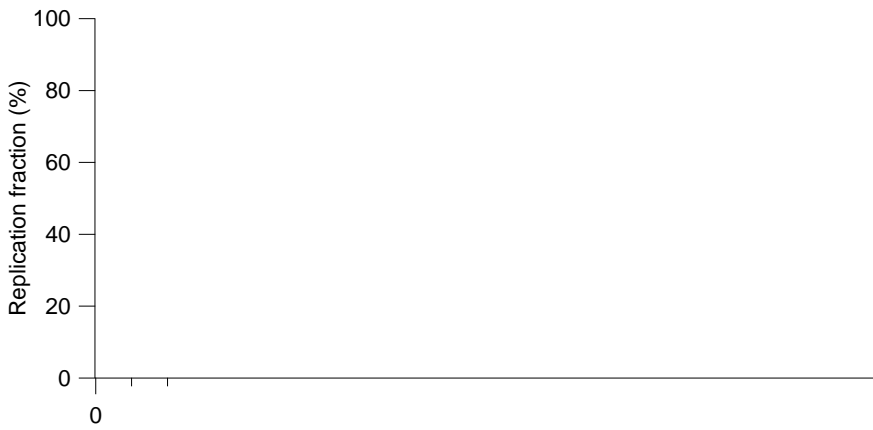
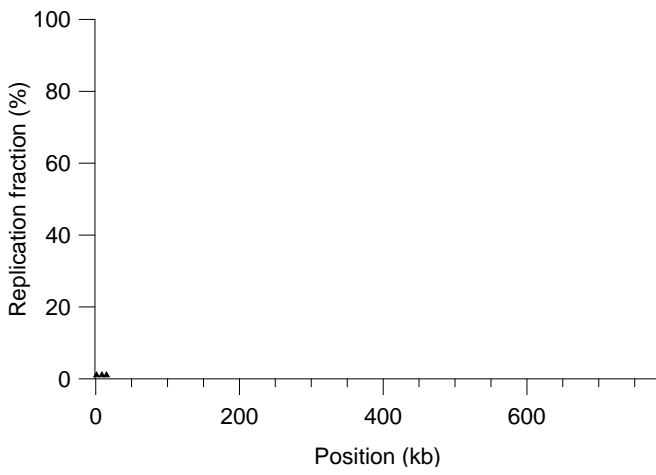
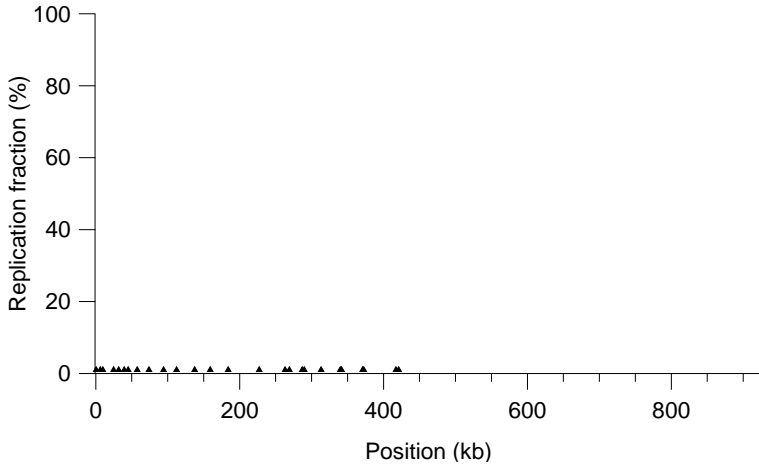


Supplementary Figures

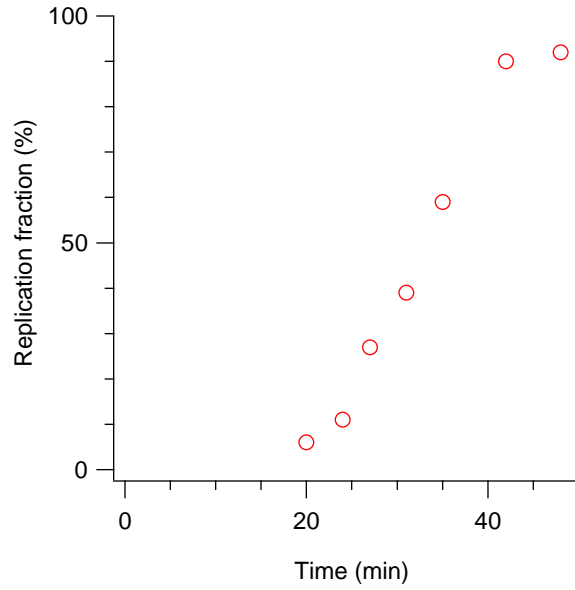






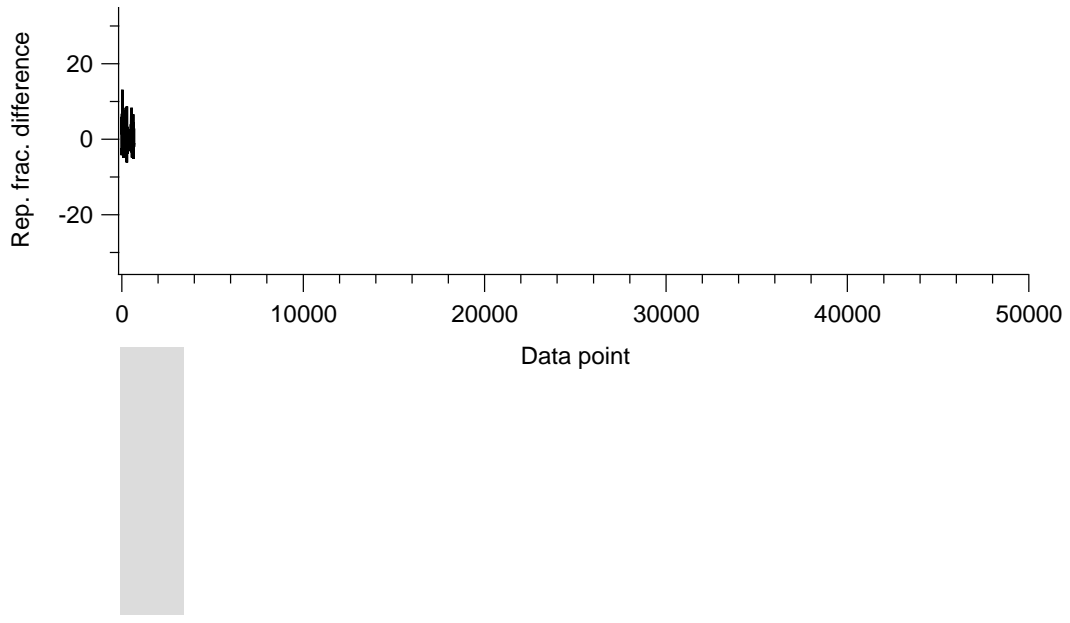


SUPP. FIG. 1: Genome-wide SM and MIM tests, separately shown for each chromosome. Roman numeral corresponds to chromosome number. The x-axis denotes the position along the chromo-



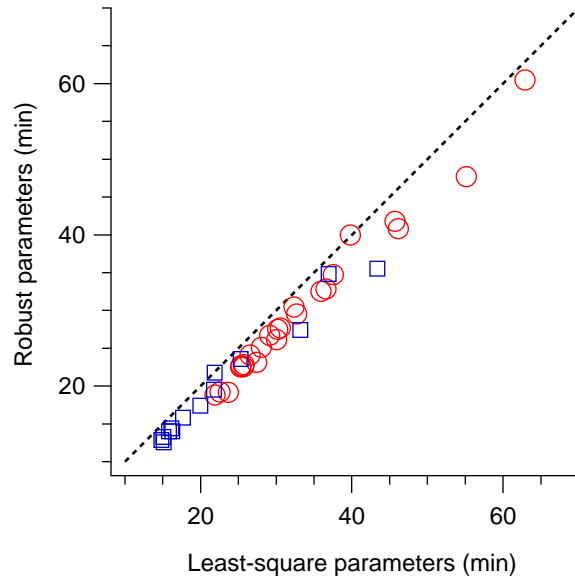
SUPP. FIG. 2: Replication fraction of ARS501. ARS501 is located on chromosome V at 549 kb. Circles are data from a slot-blot experiment (Ferguson *et al*, 1991); squares are data from the newer microarray experiment (McCune *et al*, 2008). Lines are fits to the data using a sigmoid (Hill equation). Values 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, significance level = 0.05).



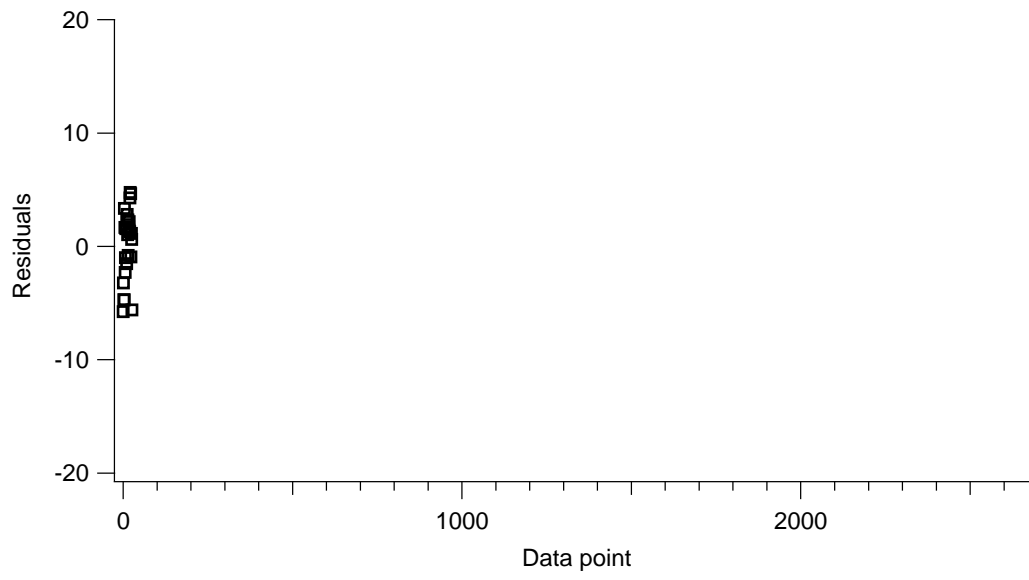
SUPP. FIG. 4:

SUPP. FIG. 4: **A.** Difference between two equivalent experiments from McCune *et al* 2008. The differences between the replication fraction of two nominally equivalent exper-



SUPP. FIG. 5: Comparison between least-squares and robust t parameters for chromosome XI. The x-axis corresponds to the least-squares t , 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.


A



B

C

5.00 μ 0



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 *cdc7-1*. 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 $N(\mu, \sigma)$ 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 ν) and correspond to the same time point. The only difference among them is the starting-time distribution. The theoretical calculation (solid curves) matches the simulations (dashed curves) well. Horizontal dashed lines are the replication fraction 0-lines for the three cases. **B.** Comparison of $t_{1=2}$ 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 parameters. The x-axis, y-axis, and dotted lines are as described in B.

A

B

C

SUPP. FIG. 9: **A**. Residuals of SM t to the smoothed data of chromosome XI. the rst 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

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 *clb₅* *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.ori.db.org>
- Niemisto A, Nykter M, Aho T, Jalovaara H, Marjanen K, Ahdesmaki 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 Bioinform 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)